

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/4321>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

Bayesian Graphical Forecasting Models for Business Time Series.

Catriona M. Queen

Thesis submitted for the degree of Doctor of Philosophy
at the University of Warwick.

Department of Statistics,
University of Warwick.

September 1991.

Contents

1	Introduction.	1
1.1	The Original Problem.	1
1.2	Outline of the Thesis.	3
2	Market Structures and Market Models.	5
2.1	Competitive Market Structures.	5
2.2	The General Structure of the Markets of Interest.	8
2.3	Ideal Properties of Competitive Market Models.	11
2.4	Multivariate Models Already Developed and Their Unsuitability for This Application.	11
3	Bayesian Forecasting and Dynamic Models.	16
3.1	Dynamic Linear Models.	17
3.2	Discount Factors.	21
3.3	DLM with Unknown Observation Variance.	22
3.4	Monitoring and Intervention.	26
3.5	Multi-process Models.	28
3.6	Dynamic Multivariate Regression Models.	30
4	An Introduction to Graphical Models.	36
4.1	Influence Diagrams.	36
4.2	Directed Markov separation.	39
4.3	Decomposable Graphs.	42
4.4	Chain Graphs.	48
4.5	Granger Causality and Conditional Independence.	51
4.6	Influence Diagram of the DLM.	52
5	Multiregression Dynamic Models.	54
5.1	Introduction.	54
5.2	Heuristic Motivation for MDM's Provided by Partially Segmented Markets.	55
5.3	The Multiregression Dynamic Model.	59
5.4	Formal presentation of the results.	63

5.5 Linear Multiregression Dynamic Models. 70

5.5.1 Example. 77

5.5.2 Example. 83

5.6 Discussion of the MDM. 85

6 Dynamic Graphical Models. 90

6.1 Introduction. 90

6.2 Dynamic Graphical Models. 93

6.3 Linear Dynamic Graphical Models. 100

6.4 A Simple Illustration of a DGM. 106

6.5 Conclusion 109

7 Partial Segmentation Models. 111

7.1 Introduction. 111

7.2 Setting up the model. 113

7.3 Model Consistent Solutions. 116

7.3.1 Example. 123

7.3.2 Example. 125

7.4 The class of simple Z matrices. 126

7.4.1 Example. 130

7.5 Representing Outcomes as Cliques on a Graph of Records. 134

7.6 D.n.h. Reparametrisation Using Graphical Results. 140

7.7 Conclusion 147

8 Discussion and further research. 149

8.1 Integration of Partial Segmentation Models and MDM's/DGM's. . 150

8.2 Other topics for Further Research. 157

A Consistency and Estimability of Partially Segmented Markets. 159

List of Figures

2.1	Heuristic undirected graph of market segments.	10
4.1	The graph of an influence diagram defined on 4 variables.	38
4.2	Illustration of directed Markov separation.	41
4.3	Two graphs of ID's, one of which is decomposable.	43
4.4	Decomposable graph J together with J^u	43
4.5	Example of a triangulated graph.	44
4.6	A complete graph.	44
4.7	Example of a graph with 4 cliques.	45
4.8	An undirected graph following the RIP.	46
4.9	An undirected graph.	47
4.10	A chain graph.	49
4.11	A chain graph with its moralised graph.	51
4.12	Influence diagram of DLM before observation y_t	53
4.13	Influence diagram of DLM after obseving y_t	53
5.1	Heuristic representation of a partially segmented market.	56
5.2	Graph of an ID of brand sales over time.	57
5.3	Graph of ID representing 4 variables to illustrate an MDM.	60
5.4	Graphs for proof of theorem.	64
5.5	Graphs for proof of theorem.	65
5.6	Graphs for proof of theorem.	66
5.7	Graph of ID to prove corollary.	69
5.8	Graph for a simple LMDM	77
5.9	Contour plots demonstrating the non-normality of variables mod- elled by MDM's	80
5.10	Contour plots demonstrating the non-normality of variables mod- elled by MDM's	81
5.11	Bivariate 3-D plots of LMDM with unknown variances.	82
5.12	Two graph with the same implied conditional independences, but with different LMDM's.	87
6.1	Graph representing the research hypothesis of the shampoo mar- ket.	91

6.2	Chain graph representing a research hypothesis of a market. . . .	93
6.3	One-step-ahead forecasts of shampoo market illustrating linear DGM's.	108
6.4	Contour plots of initial joint forecast densities of (a) $Y(1)$ and $Y_1(2)$, and (b) $Y_1(2) + Y_2(2)$ and $Y(3)$	109
7.1	Time series of parameter evolution.	133
7.2	Graph illustrating G-sets.	135
7.3	Graph of a partitioning matrix.	135
7.4	Graph of example 7.3.1.	138
7.5	Graph of example 7.3.2.	138
7.6	Graph of example 7.4.1.	139
7.7	Two graphs which are non-identifiable for θ	139
7.8	$G(Z)$ for two recursive-directed matrices.	141
7.9	Dnh reparameterisation using graph G	146
8.1	$G(Z)$ of example	151
8.2	Graph of ID of transformed variables.	153
8.3	$G(Z)$ of star model example.	154
8.4	Star model example — graph of ID	156

List of Tables

2.1	Analysed panel data.	9
2.2	Purchase probabilities in a partially segmented market.	15
7.1	Bi-weekly sales of 7 brands of computer.	130

Acknowledgements.

I would like to thank my supervisor, Jim Smith, for all his help, encouragement and guidance over the last three years. I would also like to thank the staff of the statistics department, especially Ewart Shaw who spent many hours trying to help me draw some contour plots. Thanks are also due to P.Jedeja and Jamal Ameen at Unilever Research, who have given me some insight into the problem of modelling competitive business markets and have helped me to gain some experience of analysing industrial time series.

I would especially like to thank David for not only keeping me (relatively) sane throughout this study, but also for the many, many hours spent helping me to format data and draw and print diagrams, many of which appear in this thesis.

Finally I would like to thank both the Science and Engineering Research Council and Unilever Research for their financial assistance throughout this research.

To David

Eeyore was saying to himself. "This writing business. Pencils and what-not. Over-rated, if you ask me. Silly stuff. Nothing in it."

Winnie-the-Pooh

Summary.

This thesis develops three new classes of Bayesian graphical models to forecast multivariate time series. Although these models were originally motivated by the need for flexible and tractable forecasting models appropriate for modelling competitive business markets, they are of theoretical interest in their own right.

Multiregression dynamic models are defined to preserve certain conditional independence structures over time. Although these models are typically very non-Gaussian, it is proved that they are simple to update, amenable to practical implementation and promise more efficient identification of causal structures in a time series than has been possible in the past.

Dynamic graphical models are defined for multivariate time series for which there is believed to be symmetry between certain subsets of variables and a causal driving mechanism between these subsets. They are a specific type of graphical chain model (Wermuth & Lauritzen, 1990) which are once again typically non-Gaussian. Dynamic graphical models are a combination of multiregression dynamic models and multivariate regression models (Quintana, 1985, 87, Quintana & West, 1987, 88) and as such, they inherit the simplicity of both these models.

Partial segmentation models extend the work of Dickey et al.(1987) to the study of models with latent conditional independence structures. Conjugate Bayesian analyses are developed for processes whose probability parameters are hypothesised to be dependent, using the fact that a certain likelihood separates given a matrix of likelihood ratios. It is shown how these processes can be represented by undirected graphs and how these help in its reparameterisation into conjugate form.

Chapter 1

Introduction.

This thesis was originally motivated by the practical problem of developing classes of Bayesian forecasting models appropriate for competitive business markets. Although there are aspects of this problem which still remain unresolved by this thesis, it is hoped that the models developed here create a foundation on which to base further research into this problem, as well as being of theoretical interest in their own right. This introduction firstly gives a brief summary of the original problem motivating the research and then presents an outline of the thesis.

1.1 The Original Problem.

Competitive product markets have several different brands available of the same product. Each brand in the market uses various advertising, promotion and pricing strategies — known here as “competitive strategies” — to try and retain its present share of the market or increase it. Thus, all the companies in the market are continually making decisions concerning such problems as setting the advertising and promotions budget for the coming year; setting the brand’s retail price at a competitive rate; deciding whether an advertising campaign would be preferential to a promotion at a particular time; and when, where and how to

optimally advertise/promote their brand.

The total volume of sales in a market is usually approximately constant over time, so that if one brand increases its sales, another brand must lose sales. Therefore, the competitive strategies of any particular brand will not only be expected to affect its own sales, but will also often affect the sales of competing brands. Many different situations can influence the effect that a competitive strategy can have on the various brand sales. Several brands can employ competitive strategies simultaneously, thus possibly limiting the full effects of a competitive strategy. Brands affected by a competitive strategy can be expected to retaliate by employing their own strategy designed to counter the effects of the original strategy. The sequence in which the competitive strategies and their retaliations occur and their relative timings are also important factors in influencing the effects of any strategy. Suppose, for example, that a brand has just had a generous promotion in which for every purchase of a packet of that brand, the consumer gets a packet free. Suppose that a large proportion of consumers have taken advantage of the promotion and now do not need to purchase the product again for a few months. If the first promotion continues for long enough, a promotion offering a 10% discount, for example, will seem to consumers small in comparison. Therefore, any retaliation promotion would have to seem at least as generous for consumers to compare it favourably with the previous promotion. Thus, for the retaliation to have an optimal effect, the type of retaliation chosen must take into account the type of the original strategy. The timing of the retaliation is also influenced by the previous strategy. If a brand retaliated immediately, then this would not be as profitable as it would be if the brand waited until the consumers used up their stock before retaliating. It is therefore important that a brand makes a wise decision as to when, where and how it employs its competitive strategy so that

in the face of competition the brand's strategy can have the optimal effect.

If a brand had information about the future strategies of the other brands in the market, then this knowledge could be used so that this brand could make optimal decisions to suit the competing brands' future strategies. Of course, in reality, each brand will not have information about the future strategies of competing brands, and they must use their knowledge of the competitive behaviour and its expected effects on the brand sales in that particular market to predict the future strategies. Therefore, many companies and organisations are very interested in developing an understanding of the competitive behaviour in specific product markets to try and help improve their decision making and train new decision makers in the consequences of their actions.

1.2 Outline of the Thesis.

Chapter 2 considers the different possible market structures and the general structure of the markets of interest is highlighted. The various properties which realistic forecasting models for competitive markets ideally should have are also examined and the limitations of previously developed models for this application are discussed. Chapters 3 and 4 which provide introductions to Bayesian forecasting techniques and graphical models respectively, contain the background knowledge needed before the new forecasting models developed in this thesis are introduced. In chapter 5, a new class of Bayesian forecasting model is developed which defines a conditional independence structure across the brand sales in a market and utilises any heuristic causal relationships which might exist amongst the brands. Chapter 6 extends these models so that more complex relationships can be accommodated. A new class of forecasting model based on the generalised Dirichlet distribution is defined in chapter 7 which models the brand market share. Finally

in chapter 8 it is shown how the models of chapter 7 can be generalised into a form compatible with the models of chapters 5 and 6 and some ideas for further research of the problem are considered.

Chapter 2

Market Structures and Market Models.

This chapter is essentially divided into two parts. Firstly, section 2.1 examines the various types of structure that can occur in competitive markets while special attention is paid to the structure of interest in this thesis in section 2.2. The second part of the chapter firstly examines in section 2.3 some of the desirable properties of a model for competitive markets and secondly in section 2.4 discusses the reasons why the multivariate models developed to date are not ideal for modelling competitive markets. In particular, it is shown how one of the most popular market models — the Dirichlet model — is inappropriate for this application.

2.1 Competitive Market Structures.

Suppose a market has n brands, B_1, \dots, B_n . Generally market structures can be of two types — homogeneous or heterogeneous. If it is *homogeneous*, then each potential customer has the *same* probability ψ_j of buying brand j , $j = 1, \dots, n$. On the other hand, if it is *heterogeneous*, then it is assumed that consumers can be divided into r types T_1, \dots, T_r where customers within each type are

homogeneous. It is established which brands each type T_i usually buys and these are grouped into the set of brands B_{T_i} , for $i = 1, \dots, r$. A heterogeneous market is then defined by the following conditions:

1. For each $i = 1, \dots, r$ there is some j , $1 \leq j \leq n$ for which

$$P(\text{buy } B_j \mid \text{customer type } T_i) > 0$$

so that each customer type buys at least one brand.

2. For each $j = 1, \dots, n$ and some i , $1 \leq i \leq r$

$$P(\text{buy } B_j \mid \text{customer type } T_i) > 0$$

so that each brand is bought by at least one customer type.

3. There is at least one j , $1 \leq j \leq n$ for which:

$$P(\text{buy } B_j \mid \text{customer type } T_i) \neq P(\text{buy } B_j \mid \text{customer type } T_k) \neq 0,$$

for some $i \neq k$, $1 \leq i, k \leq r$. In other words, there is at least one brand for which the purchase probabilities are not homogeneous across customers.

Note also that $\sum_j P(\text{buy } B_j \mid \text{customer type } T_i) = 1$, for each i .

Now suppose that for each customer type T_i , $1 \leq i \leq r$:

$$P(\text{buy } B_j \mid \text{type } T_i) > 0, \quad \text{for } B_j \in B_{T_i}$$

and

$$P(\text{buy } B_k \mid \text{type } T_i) = 0, \quad \text{for } B_k \in B_{T_i}^c \neq \emptyset$$

where $B_{T_i}^c$ denotes the complement of the set B_{T_i} . Notice that by 1) above:

$$B_{T_i} \neq \emptyset$$

and by 2) above

$$(B_{T_1} \cup B_{T_2} \cup \dots \cup B_{T_r}) = (B_1 \cup B_2 \cup \dots \cup B_n).$$

Now if B_{T_1}, \dots, B_{T_r} form a partition of B_1, \dots, B_n then this heterogeneous market is called a *segmented market*. However, if there are some values i, k such that for $i \neq k, 1 \leq i, k \leq r$:

$$(B_{T_i} \cap B_{T_k}) \neq \emptyset$$

then this is called a *partially segmented market*.

Smith (1956) first defined the concept of market segmentation and ever since it has played a major role in marketing. Knowing that a brand appeals to a certain type of customer can dictate where, when and how competitive strategies for that brand are employed. It is therefore very important for companies to identify the segmentation in a market so that differences between customer types can be both accounted for and utilised when marketing a product to help increase profitability (Frank et al., 1972, Wind, 1978, Samli, 1989).

The main problem concerning the concept of market segmentation is that there are many different ways in which the types of customer have been defined and each company makes a subjective choice about the segmentation in a market depending on how they define the customer types. Consumers have been segmented according to their geography; demographic and socioeconomic factors such as age, sex, education, occupation; life-style and personality differences; and the type of brand-use. However, Roberts & Docker (1986) have argued that segmenting consumers according to these qualities are not in fact good indicators of their brand choices. Instead they propose segmenting consumers according to their attitudes to brands and how they perceive them. Some studies have shown that consumers generally agree as to what attributes they believe each brand

has, only they rank the attributes differently and it is the ranking which creates the different consumer types. Bourgeois et al. (1980, 82) and Day et al. (1979) utilise this approach so that markets are segmented according to the views of actual or potential customers and those brands which can be considered as good substitutes for each other are put into the same segment.

2.2 The General Structure of the Markets of Interest.

Different modelling approaches are required depending on whether the market of interest is homogeneous, segmented or partially segmented. It is therefore important to establish the structure of a market before any model for it is derived. The markets modelled in this thesis are a selection of the markets which are of interest to Unilever Research, such as the washing powder and soap markets. Markets of the same type of product usually exhibit the same general structure. So by investigating the structure of one of the markets of interest, forecasting models designed for the specified market structure could be developed.

Consumer panel data for a particular market was analysed to establish whether the market was homogeneous, segmented or partially segmented. The analysed market consisted of 24 brands, B_1, \dots, B_{24} . The data listed the brand purchases of 3678 households over a 52 week period. As very little information was available about the possible customer types for the market, the consumers were partitioned into types T_1, \dots, T_{24} depending on which brand they bought last. That is, T_i consists of all those consumers who bought brand i as their last purchase. The market was dominated by just 8 brands labelled B_1, \dots, B_8 . The market shares of the remaining brands were so small that these brands were amalgamated to form a single brand B_{rest} . Similarly, T_{rest} consists of all the consumers whose last purchase was any brand in B_{rest} . The total number of purchases made of brands

BRANDS

CUST. TYPE		B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_{rest}
	T_1	.308	.182	.150	.097	.074	.028	.028	.017	.115
	T_2	.311	.209	.125	.099	.079	.028	.022	.027	.097
	T_3	.295	.183	.167	.109	.085	.026	.023	.025	.084
	T_4	.298	.223	.139	.106	.060	.023	.029	.031	.091
	T_5	.354	.186	.118	.095	.094	.014	.018	.028	.092
	T_6	.407	.195	.137	.050	.073	.074	.014	.011	.039
	T_7	.391	.211	.073	.050	.047	.032	.074	.004	.118
	T_8	.225	.223	.142	.087	.031	.030	.014	.132	.115
	T_{rest}	.272	.207	.129	.103	.073	.038	.041	.030	.108

Table 2.1: Proportions of each brand bought by each consumer type.

B_1, \dots, B_8 and B_{rest} by each of the types T_1, \dots, T_8 and T_{rest} was calculated. The proportions of each brand bought by each customer type is displayed in the table 2.1.

Now if the market is homogeneous, then it is expected that the probability that brand j is purchased by any customer is ψ_j . Further, it is assumed that successive purchases are independent and that ψ_j remains constant over time. Therefore, the consumers in each different customer group should be buying the various brands in the same proportions. However, when a χ^2 test of homogeneity was carried out, it was found that the proportions of each brand j bought were not homogeneous across the customer type. Thus, as this market is a typical example of the markets of interest, this thesis concentrates on developing forecasting models for heterogeneous markets.

As was mentioned at the end of section 2.1 there are many different ways of defining the segmentation in the market depending on how the customer types are classified. In this thesis, the customer types are defined by using a similar approach to that of Bourgeois et al. (1980, 82) and Day et al. (1979). Suppose that customers are divided into r types T_1, \dots, T_r depending on which attributes,

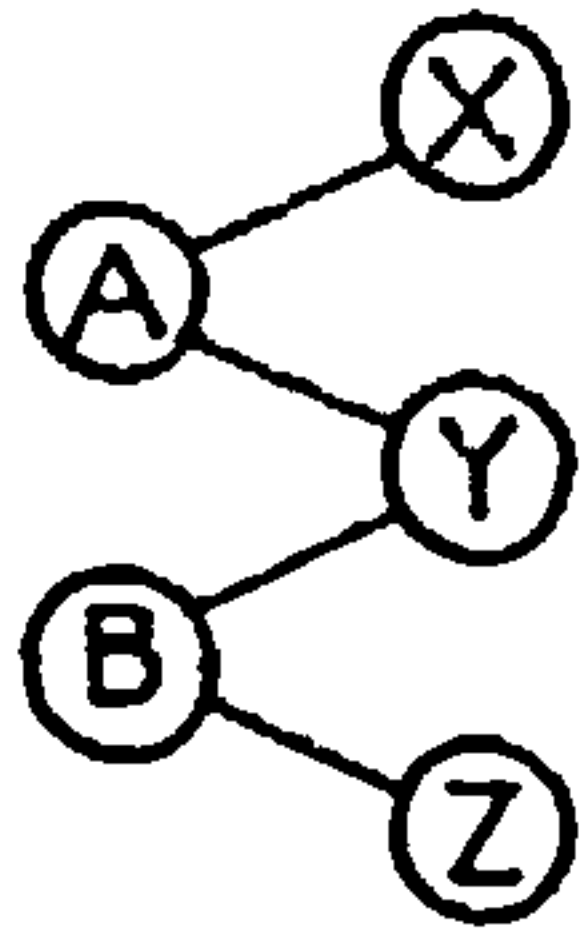


Figure 2.1: Heuristic undirected graph of market segments.

such as price or quality, that they consider most important in a brand. The brands in the market which consumers believe possess the attribute specified by T_i are put collectively into the subset B_{T_i} . Thus if any brand has more than one set of attributes and therefore appeals to more than one customer type, say T_i and T_k , then $(B_{T_i} \cap B_{T_k}) \neq \emptyset$ and so this will form a partially segmented market.

Consider the simplest possible partially segmented market. Suppose that any consumer has a choice of just three brands X , Y and Z in a product market and that

- X has attribute A alone
- Y has both attributes A and B
- Z has attribute B alone

Now this market can be divided into two segments using the two attributes A and B . This can be represented heuristically by the undirected graph (see chapter 4) in figure 2.1.

This heuristic graphical interpretation of partially segmented markets provides the initial motivation to amalgamate graphical models (see chapter 4) with Bayesian forecasting models (see chapter 3) to produce the new classes of Bayesian forecasting models for competitive markets introduced in chapters 5 and 6, as well as providing a useful basis for a pictorial representation of the models introduced in chapter 7.

2.3 Ideal Properties of Competitive Market Models.

Apart from requiring multivariate models which can accommodate the market structures introduced in section 2.2, there are other features which ideally any mathematical model for competitive markets should have.

Firstly, the effects of competitive policies on various brand sales tend to be non-linear with time (Migon & Harrison, 1985, Colman & Brown, 1983). Thus any realistic model needs to be non-linear. The effects of competitive policies do not necessarily remain static with time. For example, suppose that a brand starts a three month major advertising campaign. Clearly the initial effect of the campaign on that brand's sales will not be the same as during the following three months. Therefore, any realistic model needs to be dynamic and allow the parameters to drift with time. Finally, because these models are of great interest to companies, they need to be simple enough to allow for practical implementation.

2.4 Multivariate Models Already Developed and Their Unsuitability for This Application.

Multivariate time series have been analysed extensively in recent years. However, none of the models developed so far are appropriate for competitive markets because they do not satisfy all the criteria outlined in section 2.3 which are desirable for this application.

Most of the attention has focussed, implicitly or explicitly, on the study of linear, time homogeneous processes where large amounts of data are available to produce relatively complex models (Hannan & Kavalieris, 1984, Robinson, 1973, Jewell & Bloomfield, 1983, Jewell et al., 1983). Economists also tend to restrict their attention to linear, time homogeneous time series models. However,

their models are more amenable to practical implementation as they have a more structured approach. They limit their attention to simple, but plausible, models which they reject only with strong evidence from the data (Chan & Wallis, 1978, Engle & Granger, 1987, Harvey, 1989, Harvey & Stock, 1988, Stock & Watson, 1988). The Bayesian linear models of Quintana (1985) and Harvey (1986) are also relatively simple to implement and, in addition, do not assume time homogeneity of the series allowing the parameters to drift with time. However, they require stringent symmetry between the component series and so do not accommodate the complex relationships which can exist between the brand sales in competitive markets.

Goodhardt et al. (1984) developed a model of consumer buyer behaviour called the *Dirichlet model*. The model considers the number of purchases made of each brand by a single particular consumer i and then generalises this to find a model for the number of brand purchases of any consumer. The Dirichlet model will be discussed in a little more detail here as it has been considered as one of the most useful market models to date.

Suppose there are n brands in a product market. The Dirichlet model is defined as follows.

1. The i^{th} individual, $i = 1, 2, \dots$, has a probability $0 \leq (\psi_j)_i \leq 1$ of choosing brand j from the n alternatives. It is assumed that these probabilities are fixed over time and that successive purchases are independent. The number of purchases, $(r_j)_i$, that customer i makes of brand j , $j = 1, \dots, n$, in any one time period follows a *multinomial distribution* so that for the i^{th} individual:

$$p((r_1)_i, \dots, (r_n)_i | (\psi)_i) = \frac{N_i!}{\prod_{j=1}^n (r_j)_i!} \prod_{j=1}^n (\psi_j)_i^{(r_j)_i}$$

where $\sum_{j=1}^n (r_j)_i = N_i$.

2. The brand choice probabilities ψ_j vary across all consumers according to a *Dirichlet distribution*. So for any general consumer:

$$p(\psi_1, \dots, \psi_n | \alpha) = \frac{\Gamma(\sum_{j=1}^n \alpha_j)}{\prod_{j=1}^n \Gamma(\alpha_j)} \prod_{j=1}^n \psi_j^{\alpha_j - 1} \quad (2.1)$$

where $\sum_{j=1}^n \psi_j = 1$.

3. In any one time period the total number of purchases made by the i^{th} individual follows a *Poisson process* with rate $(\mu)_i > 0$ so that:

$$p(N_i | (\mu)_i) = \frac{(\mu)_i^{N_i} e^{-(\mu)_i}}{N_i!}.$$

4. A *gamma distribution* describes how the mean purchasing rates vary amongst individuals so that for a general consumer:

$$p(\mu | \beta, \delta) = \frac{\delta^\beta}{\Gamma(\beta)} \mu^{\beta-1} e^{-\delta\mu}.$$

Therefore, for any consumer, the probability that they make N purchases in total is given by:

$$p(N | \mu) = \frac{\mu^N e^{-\mu}}{N!}.$$

So, for any consumer, the probability that in any one time period they buy r_j of brand j , for $j = 1 \dots, n$, is given by:

$$p(r_1, \dots, r_n | \psi, N) = \frac{N!}{\prod_{j=1}^n r_j!} \prod_{j=1}^n \psi_j^{r_j}. \quad (2.2)$$

The Dirichlet model is relatively simple to use, can give reasonable forecasts and, through a closed form analysis, could provide the basis of a Bayesian forecasting model allowing the parameters to change with time. However, unfortunately it has been designed for homogeneous markets and is inappropriate for partially segmented markets. This is because one of the properties of the Dirichlet distribution is not compatible with partially segmented markets. This result will now be shown in more detail.

Let $\psi = (\psi_1, \dots, \psi_n)$ follow a Dirichlet distribution such that:

$$p(\psi | \alpha) \propto \psi_1^{\alpha_1} \psi_2^{\alpha_2} \dots \psi_n^{\alpha_n}, \quad \psi_j \geq 0, \quad \sum_{j=1}^n \psi_j = 1$$

so that if $\psi_1 = 0$ then:

$$p(\psi | \psi_1 = 0, \alpha) \propto \psi_2^{\alpha_2} \dots \psi_n^{\alpha_n}, \quad \psi_j \geq 0, \quad \sum_{j=2}^n \psi_j = 1.$$

So in particular

$$\frac{E(\psi_j | \psi_1 = 0)}{E(\psi_i | \psi_1 = 0)} = \frac{E(\psi_j)}{E(\psi_i)} = \frac{\alpha_j}{\alpha_i}, \quad \text{for } i, j \neq 1. \quad (2.3)$$

Now suppose there is a partially segmented market with three brands B_1, B_2, B_3 and two customer types T_1 and T_2 . Suppose that

$$P(\text{buy } B_j) = \psi_j, \quad j = 1, 2, 3,$$

$\psi_j \geq 0, \sum_{j=1}^3 \psi_j = 1$ and that

$$P(\text{any purchase bought by type } T_i) = \theta(i), \quad i = 1, 2,$$

$\theta(i) > 0, \theta(1) + \theta(2) = 1$. Table 2.2 defines a partially segmented market for some p and $q > 0$. Each entry is the probability that brand B_j is purchased given that type T_i is buying.

		BRANDS		
CUST. TYPE		B_1	B_2	B_3
	T_1	p	$1 - p$	0
	T_2	0	$1 - q$	q

Table 2.2: Purchase probabilities in a partially segmented market where each entry is $P(\text{Buy } B_j | \text{type } T_i)$, $j = 1, 2, 3$, $i = 1, 2$.

If (p, q) have any (non-degenerate) distributions defined on them, then:

$$E(\psi_2) = \theta(1)E(1 - p) + \theta(2)E(1 - q)$$

$$E(\psi_3) = \theta(2)E(q)$$

but

$$E(\psi_2 | \psi_1 = 0) = \theta(1) + \theta(2)E(1 - q)$$

$$E(\psi_3 | \psi_1 = 0) = \theta(2)E(q).$$

Therefore

$$\frac{E(\psi_2 | \psi_1 = 0)}{E(\psi_3 | \psi_1 = 0)} = \frac{E(\psi_2)}{E(\psi_3)} + C, \quad \text{where } C = \frac{\theta(1)E(p)}{\theta(2)E(q)} > 0.$$

As property 2.3 of a Dirichlet distribution does not hold across $\{\psi_1, \psi_2, \psi_3\}$, a Dirichlet model cannot be used on partially segmented markets such as this one. However, chapter 7 defines a forecasting model which is a generalisation of the Dirichlet model. This basically models each segment of the market by a Dirichlet model and has extra model parameters representing segment intersections.

Chapter 3

Bayesian Forecasting and Dynamic Models.

A scientific model is a description of a system. A time series model is a description of the past, present and future values of a time series of observations. A forecast is a conjecture about something in the future. Time series models can be used as a means of both learning about the system and also making forecasts about future observations.

It is often assumed that one wants to learn about a system to find a “true” model. However, it has been argued that no model can be a “true” representation of a system (Maybeck, 1979, 1982). Harrison & Stevens (1976) developed a Bayesian approach to forecasting where a model is a description of the system *as perceived by the modeller*. It is this approach to time series modelling which is used in this thesis.

Dynamic Bayesian forecasting models are a class of probabilistic models which represent a subjective view of a system. Initially, when little or no data is available, the forecaster uses expert or subjective knowledge to set up their model. The model then evolves sequentially with each observation of the series. The forecaster may intervene into the forecasting system whenever their beliefs about the

model change or external events suggest that the model is no longer appropriate. All forecasts are represented in terms of probability distributions, thus reflecting the amount of uncertainty which exists as a result of using subjective models. Monitoring techniques are used both to check that these subjective models are satisfactory and also to detect any unanticipated major changes in the series.

This chapter gives a short introduction to linear dynamic Bayesian forecasting. Section 3.1 gives a brief introduction to the general form of the most well-known Bayesian forecasting model, called the dynamic linear model, which is defined here for a multivariate time series. The following few sections introduce some of the techniques initially developed for this model, which are either mentioned later in this thesis, or can be directly applied to the new classes models introduced in chapters 5, 6 and 7. Models in which the forecaster need not specify fixed values of the variances are discussed in sections 3.2 and 3.3, while monitoring techniques together with the use of subjective intervention are introduced in section 3.4 and multi-process models are discussed in section 3.5. All these techniques are described in terms of the univariate dynamic linear model as this gives a simple presentation of the basic ideas. Finally in section 3.6, a multivariate extension of the dynamic linear model introduced in section 3.3 is presented, which allows the variance matrix of the series to be estimated on-line as the series develops.

3.1 Dynamic Linear Models.

Let the n -dimensional multivariate time series at time t be denoted by the column vector Y_t such that:

$$Y_t^T = (Y_{t1}, \dots, Y_{tn})$$

where X^T denotes the transpose of a matrix X . If y_{ij} is the observed value of Y_{ij} , then the conventional notation will be used that $y_j^t = (y_{1j}, \dots, y_{ij})$ and similarly $Y_j^t = (Y_{1j}, \dots, Y_{ij})$.

The *normal dynamic linear model* or, when normality is assumed (West & Harrison, 1989a, p105), the *dynamic linear model* (DLM) (Harrison & Stevens, 1976) defines a model for Y_t in terms of an additional s -dimensional vector time series of *states* $\{\theta_t\}_{t \geq 1}$. A DLM is defined by an *observation equation*, which is a model for Y_t in terms of the state vectors; a *system equation*, which relates the state vectors at time t with those at time $t - 1$; and the *initial information* about the system which is represented through a probability distribution of the state vector prior to any observations of the series. Let the knowledge available to the forecaster at time 0 be represented by D_0 . The DLM for Y_t is then given by:

Observation equation

$$Y_t = F_t^T \theta_t + v_t, \quad v_t \sim N(\mathbf{o}, V_t) \quad (3.1)$$

System equation

$$\theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim N(\mathbf{o}, W_t) \quad (3.2)$$

Initial information

$$(\theta_0 | D_0) \sim N(m_0, C_0) \quad (3.3)$$

where F_t is a known $(s \times n)$ matrix of independent variables; v_t is the n -dimensional *observation error* vector; V_t is a known $(n \times n)$ observational variance matrix; G_t is a known $(s \times s)$ *evolution matrix*; w_t is the s -dimensional *system error* vector; W_t is a known $(s \times s)$ evolution variance matrix; and $N(\mu, \Sigma)$ represents the multivariate normal distribution with mean μ and variance Σ . It is assumed that v_t , w_t and θ_{t-1} are mutually independent where the sequences

$\{v_t\}_{t \geq 1}$ and $\{w_t\}_{t \geq 1}$ are also independent with time. Initially the forecaster must specify values for m_0 and C_0 and the series V_t and W_t .

By definition, the distribution of Y_t is completely specified by θ_t , so the joint forecast distribution of the vector time series at any time t , is given by :

$$p(y_t|y^{t-1}) = \int_{\theta_t} p(y_t|\theta_t)p(\theta_t|y^{t-1})d\theta_t. \quad (3.4)$$

If F_t , G_t , V_t and W_t are not time dependent, then the DLM is known as a *constant* DLM. If $W_t = 0$ also, then this is known as a *noise-free constant* DLM. Many DLM's can be quite complex. In this case they are generally constructed by using the *superposition* principle in which the final DLM is built from simpler components representing the level of the process, the trend, the seasonality, etc.

After each set of observations y_t , the beliefs about the state vector (and thus the future forecasts of the series) are updated. The derivation of the updating distributions come directly from multivariate normal distribution theory (see West & Harrison, 1989a, p599-600 and 111-113 for details). As the same updating process occurs at each time period, only the updating process from beliefs at time $t - 1$ to time t will be shown here. At time $t - 1$, after observing y^{t-1} , the posterior beliefs about θ_{t-1} can be represented by the probability distribution:

$$\theta_{t-1}|y^{t-1} \sim N(m_{t-1}, C_{t-1})$$

with some mean m_{t-1} and some variance C_{t-1} . Through the system equation this posterior distribution leads straight to the prior distribution:

$$\theta_t|y^{t-1} \sim N(a_t, R_t)$$

where

$$a_t = G_t m_{t-1} \quad \text{and} \quad R_t = G_t C_{t-1} G_t^T + W_t$$

which in turn leads easily to the 1-step ahead forecast distribution at time t :

$$Y_t|y^{t-1} \sim N(f_t, Q_t) \quad (3.5)$$

where

$$f_t = F_t^T a_t \quad \text{and} \quad Q_t = F_t^T R_t F_t + V_t. \quad (3.6)$$

Once y_t has been observed, then the distribution of θ_t can be updated to give the posterior distribution:

$$\theta_t|y^t \sim N(m_t, C_t),$$

where

$$\begin{aligned} m_t &= a_t + A_t e_t, & C_t &= R_t - A_t Q_t A_t^T, \\ e_t &= y_t - f_t & \text{and} & A_t &= R_t F_t Q_t^{-1}. \end{aligned}$$

Thus the whole cycle begins again.

Note that these recurrence relationships for updating the state vectors are essentially the same as the Kalman filter equations (Kalman 1960, 63), based on work in engineering control. Also note that the statistics m_t and C_t are sufficient for θ_t and they contain all the past history of the series required so that a forecast can be made.

Some of the established non-Bayesian forecasting techniques can be considered as special cases of the DLM. Now, as $t \rightarrow \infty$, as long as V/W , V , W , C , F and G are constant and there is observability, the adaptive coefficient A_t tends to a constant A , where every element A_{ij} in the $(s \times n)$ matrix A is such that $0 < A_{ij} < 1$, $i = 1, \dots, s$, $j = 1, \dots, n$. Thus as $t \rightarrow \infty$ the forecast mean a_t can be expressed by:

$$a_t = a_{t-1} + A(y_t - f_t).$$

For the simple case of the univariate series Y_t with steady DLM so that $n = s = 1$ and $F_t = G_t = 1$ in equations 3.1, 3.2 and 3.3, the point forecast at time t , as

$t \rightarrow \infty$, is given by:

$$\begin{aligned} m_t &= m_{t-1} + A(y_t - m_{t-1}) \\ &= (1 - A)m_{t-1} + Ay_t. \end{aligned} \tag{3.7}$$

It is straightforward to show (see West & Harrison, 1989a, p54–56) that the different forecasting techniques of Holt’s (1957) exponentially weighted moving average (EWMA), Brown’s (1962) exponentially weighted regression (EWR) and the ARIMA (0,1,1) model (Box & Jenkins, 1976), all produce the same point forecast as the asymptotic form of m_t given in equation 3.7.

3.2 Discount Factors.

The forecast performance of a model is sensitive to choosing appropriate values for both the observation error variance V_t and the evolution error variance W_t . W_t is often not particularly easy to estimate as it can be very small in comparison to the magnitude of the observations. This section introduces the concept of *discounting* which offers a relatively simple way of estimating the evolution error variance. Brown (1959, 62) first promoted the use of discounting techniques. Since then Ameen & Harrison (1984, 85a, 85b) have integrated the technique fully into DLM’s.

Discounting uses a *discount factor* which can be thought of as an indication of how durable the current quantified model is over time and how much attention is paid to historical data compared with the current observation. Explicitly, a discount factor relates C_{t-1} , the posterior variance of $\theta_{t-1} | y^{t-1}$, to R_t , the prior variance of $\theta_t | y^{t-1}$.

For ease of presentation consider the univariate steady model given in the previous section, where $n = s = 1$ and $F_t = G_t = 1$ in equations 3.1, 3.2 and 3.3.

The idea is that for a given discount factor δ where $0 < \delta < 1$:

$$W_t = C_{t-1}(\delta^{-1} - 1)$$

so that

$$R_t = C_{t-1} + W_t = \frac{C_{t-1}}{\delta}.$$

Thus W_t is simply considered as some fixed proportion of the posterior variance C_{t-1} and so it is easier to quantify δ , which represents the amount of reliance on the past, rather than quantify the evolution variance.

This concept can be applied directly to models where $s > 1$ so that:

$$R_t = \frac{G_t C_{t-1} G_t^T}{\delta}$$

where

$$W_t = G_t C_{t-1} G_t^T (\delta^{-1} - 1).$$

The concept can be extended further so that each component in the vector θ_t can have a separate discount factor reflecting the differing rates of evolution which can occur for the various parameters in the model.

3.3 DLM with Unknown Observation Variance.

As mentioned in section 3.2 the choice of the observation variance V_t , which is often difficult to estimate, is important to the forecast performance of the model. In this section a procedure for variance learning in a univariate DLM is introduced (West, 1982, Smith & West, 1983) where it is assumed that $V_t = V$ is constant but unknown. If a particular structure is imposed on the error variance sequence W_t and the prior mean m_0 , then a conjugate sequential updating procedure for V_t is available in addition to that of θ_t .

Suppose that the observation variance V_t for a univariate time series Y_t is the unknown constant V . Assuming that W_t is known, then the DLM for this series which will estimate this unknown variance is given by the following equations.

Observation equation

$$Y_t = F_t^T \theta_t + v_t, \quad v_t \sim N(0, V)$$

System equation

$$\theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim N(\mathbf{o}, VW_t)$$

Initial information

$$(\theta_0 | D_0, V) \sim N(m_0, VC_0)$$

$$(V^{-1} | D_0) \sim G(n_0/2, d_0/2)$$

where $G(\alpha, \beta)$ denotes the Gamma distribution with parameters α and β . Notice how conditional on the value of V this is essentially the same as the DLM defined by equations 3.1, 3.2 and 3.3, only the variance of $(\theta_0 | D_0, V)$ and the evolution variance are given as a multiple of the unknown observation variance V . The unknown variance V is said to have the Gamma distribution such that:

$$p(V^{-1} | D_0) = \frac{(d_0/2)^{n_0/2}}{\Gamma(n_0/2)} (V^{-1})^{(n_0/2-1)} e^{-\left(\frac{d_0}{2}\right)V^{-1}}.$$

Initially the forecaster must specify values for the quantities m_0 , C_0 , n_0 , d_0 and also the series W_t . The parameter d_0 is set so that $d_0 = n_0 S_0$ such that S_0 is the initial point estimate of V .

The updating distributions for this DLM will now be presented (for proof see West & Harrison, 1989a, p119-120.)

At time $t - 1$, the posterior distribution of θ_{t-1} , conditional on V , and the

distribution of V itself are given by:

$$\begin{aligned}(\theta_{t-1} | y^{t-1}, V) &\sim N(m_{t-1}, VC_{t-1}) \\(V^{-1} | y^{t-1}) &\sim G\left(\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right)\end{aligned}$$

for some values m_{t-1} , C_{t-1} , n_{t-1} and d_{t-1} such that $d_{t-1} = n_{t-1}S_{t-1}$ where S_{t-1} is a point estimate of V at time $t-1$. Unconditionally on V the posterior distribution of θ_{t-1} becomes:

$$(\theta_{t-1} | y^{t-1}) \sim T_{n_{t-1}}(m_{t-1}, S_{t-1}C_{t-1}).$$

This is known as a multivariate T distribution with n_{t-1} degrees of freedom, mode m_{t-1} and scale matrix $S_{t-1}C_{t-1}$. Its density is given by:

$$p(\theta_{t-1} | y^{t-1}) \propto \left[n_{t-1} + (\theta_{t-1} - m_{t-1})^T (S_{t-1}C_{t-1})^{-1} (\theta_{t-1} - m_{t-1}) \right]^{-(s+n_{t-1})/2}.$$

As the degrees of freedom $n_{t-1} \rightarrow \infty$, the distribution tends to normality so that $(\theta_{t-1} | y^{t-1})$ then has a $N(m_{t-1}, S_{t-1}C_{t-1})$ distribution.

The prior distribution of θ_t , conditional on V , is such that:

$$(\theta_t | y^{t-1}, V) \sim N(a_t, VR_t)$$

where

$$a_t = G_t m_{t-1} \quad \text{and} \quad R_t = G_t C_{t-1} G_t^T + W_t$$

with unconditional distribution:

$$(\theta_t | y^{t-1}) \sim T_{n_{t-1}}(a_t, S_{t-1}R_t)$$

The prior distribution for Y_t can be derived from this to give the conditional forecast distribution:

$$(Y_t | y^{t-1}, V) \sim N(f_t, VQ_t) \tag{3.8}$$

where

$$f_t = F_t^T a_t \quad \text{and} \quad Q_t = 1 + F_t^T R_t F_t$$

with the unconditional distribution:

$$(Y_t | y^{t-1}) \sim T_{n_{t-1}}(f_t, S_{t-1} Q_t).$$

After observing y_t , both the distributions of θ_t and V can be updated. Conditionally on V , the posterior distribution of θ_t is given by:

$$(\theta_t | y^t, V) \sim N(m_t, V C_t)$$

where

$$\begin{aligned} m_t &= a_t + A_t e_t, & C_t &= R_t - A_t A_t^T Q_t, \\ e_t &= y_t - f_t & \text{and} \quad A_t &= R_t F_t / Q_t. \end{aligned}$$

The posterior distribution of V is such that:

$$(V^{-1} | y^t) \sim G\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$$

where

$$n_t = n_{t-1} + 1, \quad d_t = n_t S_t \quad \text{and} \quad S_t = n_t^{-1} (n_{t-1} S_{t-1} + e_t^2 / Q_t).$$

Unconditionally on V the posterior distribution of θ_t then becomes:

$$(\theta_t | y^t) \sim T_{n_t}(m_t, S_t C_t).$$

Note that the prior/posterior distributions are essentially identical to those given in section 3.1, only with the observation variance V or its current point estimate, appearing in the variance term.

3.4 Monitoring and Intervention.

The Bayesian forecasting model is a subjective view of the system. Monitoring techniques (West, 1986, West & Harrison, 1986) are employed to ensure that the model is making satisfactory forecasts and to detect any unanticipated changes. Typically the forecaster would want a monitoring system to detect when the model makes serious over or under estimates of either the forecasts themselves or the variability of the series.

The basic monitoring approach discussed here is that of the *sequential probability ratio test* (SPRT) which is essentially a sequence of hypothesis tests. The idea of the SPRT is to compare the given model M with an alternative model A , where A represents a model which would signal that model M was performing badly. A forecaster will want a monitoring technique which will be able to distinguish between an outlier observation and a longer term change in the system. Thus, unless the forecast is particularly bad, the SPRT only gives a monitor signal if the forecasting system starts to give consistently poor forecasts. The forecast distribution at time t of the model M is given by $p(\cdot | \mathbf{y}^{t-1}, M)$ and that of model A is given by $p(\cdot | \mathbf{y}^{t-1}, A)$. After observing \mathbf{y}_t , the relative support for these two models is expressed through the *Bayes factor* H_t , at time t , which is given by:

$$H_t = \frac{p(\mathbf{y}_t | \mathbf{y}^{t-1}, M)}{p(\mathbf{y}_t | \mathbf{y}^{t-1}, A)}.$$

Now, at time t , one of three possible decisions is made depending on the value of H_t :

1. If $H_t > d_1$ (d_1 usually taken to be 1), then model M is considered acceptable.

2. If $H_t < d_0$ (d_0 often taken to be about 0.1), then this is a monitor signal and model M is no longer considered acceptable.
3. If $d_0 < H_t < d_1$, then the test is repeated at time $t + 1$ but this time the value of $H_t H_{t+1}$ is considered. The testing continues at each time point until the inequality $d_0 < \prod_{i=0}^T H_{t+i} < d_1$ is broken at time $t + T$ where one of two decisions can be made:
 - (a) If $\prod_{i=0}^T H_{t+i} < d_0$, then this is a monitor signal and model M is no longer considered acceptable.
 - (b) If $\prod_{i=0}^T H_{t+i} > d_1$, then model M is acceptable and the test restarts at time $t + T + 1$ and the Bayes factor H_{t+T+1} is calculated.

A forecaster may, however, have information about some event to occur in the future which they believe will affect their forecasting model. For example, a major market competitor may launch a large scale promotion which is expected to dramatically affect the sales of the forecaster's own brand. In this case, the forecaster can intervene into the forecasting system and use their subjective view to change the forecasting model to reflect the anticipated event (West & Harrison, 1989b).

A lot of uncertainty will exist concerning any possible changes in level, trend, seasonality etc. and the time it might take for these changes to occur. Despite this uncertainty, it will be important for the forecaster to have a model which can reflect this change as accurately as possible. To do this, a model must be easily adaptable to change as the history of the series becomes less relevant. The variance in the model must also be increased to reflect the added uncertainty which exists as the event occurs and to allow the model to adapt quickly. If a

sustained component change is expected then the prior distribution of the states is affected, whereas a transient change is reflected through the distribution of Y_t . The prior variances of the model control the adaption time of the model. An increase in Q_t (the forecast variance of Y_t) simply increases the variances at the change point, indicating a rapid component change. On the other hand, an increase in R_t (the forecast variance of the θ_t) allows the model to gradually adapt to new component changes. Thus intervention in a Bayesian model can be implemented so that the nature of the expected change is reflected.

3.5 Multi-process Models.

It is often unsatisfactory to describe a process in terms of just a single DLM. *Multi-process models* (Harrison & Stevens, 1976) consider a set of n alternative models $\{M^{(1)}, \dots, M^{(n)}\}$ for the series simultaneously. There are 2 classes of multi-process models, known as Class I and Class II. Class I models assume that there is a single (unknown) model $M^{(i)} \in \{M^{(1)}, \dots, M^{(n)}\}$ which is an adequate representation of the process for all time. Class II models, on the other hand, assume that no single model describes the series for all time, but rather the model considered the most appropriate from the set of possible models $\{M^{(1)}, \dots, M^{(n)}\}$ changes with time. Class II models are therefore particularly useful for modelling major changes in time series.

Multi-process models can serve one of two purposes:

1. They can form a mixture model with components $\{M^{(1)}, \dots, M^{(n)}\}$ utilising information based on the whole set of alternative models. This allows a richer model than can be provided by any single DLM alone.
2. They can be used to select the most appropriate model $M^{(i)}$ for the process

from the set of alternatives $\{M^{(1)}, \dots, M^{(n)}\}$.

A brief outline of the mixture models for both Class I and Class II multi-process models will now be presented.

As mentioned earlier, Class I models assume that a single model from the set of alternatives is an adequate representation of the process for all time. Now, after observing \mathbf{y}^{t-1} the probability that $M^{(i)}$ is the “true” model is given by:

$$p(M^{(i)} | \mathbf{y}^{t-1}) = P_{t-1}^{(i)}, \quad i = 1, \dots, n.$$

The forecast distribution for each model $M^{(i)}$, $i = 1, \dots, n$ is given by:

$$(Y_t | M^{(i)}, \mathbf{y}^{t-1}) \sim N(f_t^{(i)}, Q_t^{(i)})$$

and the forecast distribution of Y_t is a mixture of normal distributions:

$$(Y_t | \mathbf{y}^{t-1}) \sim \sum_{i=1}^n P_{t-1}^{(i)} N(f_t^{(i)}, Q_t^{(i)}).$$

After observing y_t each component DLM is updated as a single univariate DLM as shown in section 3.1. The posterior probability of $M^{(i)}$ is simply:

$$P_t^{(i)} = c f(y_t | M^{(i)}, \mathbf{y}^{t-1}) P_{t-1}^{(i)}, \quad i = 1, \dots, n$$

where c is the normalising constant.

Class II multi-process models are used when the most appropriate of the models $\{M^{(1)}, \dots, M^{(n)}\}$ varies with time. An additional n -square transition matrix $[\pi_{ij,t}]$ is defined where $\pi_{ij,t}$ denotes the probability that if model $M^{(i)}$ is obtained at time $t-1$, then model $M^{(j)}$ is obtained at time t . Alternatively, there is the simpler case where the probability that $M^{(j)}$ is obtained at any time t is π_j , regardless of the previous model. In this case the forecast distribution for Y_t is a mixture of n^2 normal distributions:

$$(Y_t | \mathbf{y}^{t-1}) \sim \sum_{j=1}^n \sum_{i=1}^n P_{t-1}^{(i)} \pi_j N(f_t^{(ij)}, Q_t^{(ij)})$$

where

$$P(M_{t-1}^{(i)}, M_t^{(j)} | \mathbf{y}^{t-1}) = P_{t-1}^{(i)} \pi_j$$

$$(Y_t | M_{t-1}^{(i)}, M_t^{(j)}, \mathbf{y}^{t-1}) \sim N(f_t^{(ij)}, Q_t^{(ij)}).$$

If multi-process models are being used so that the best model for the series can be chosen, each DLM is developed separately over time and some decision rule to choose the most suitable model is applied which compares the forecast performance of the various models. For example, if there are two possible models $M^{(1)}$ and $M^{(2)}$, then the Bayes factor of the 2 models can be used as a basis for making a decision.

3.6 Dynamic Multivariate Regression Models.

The general multivariate DLM was introduced in section 3.1. Although the discounting techniques of section 3.2 can be extended naturally to the multivariate case, the multivariate analogue when the observation covariance is unknown cannot in general follow a simple conjugate analysis. *Dynamic multivariate regression* models (DMR) models (Quintana, 1985, 87, Quintana & West, 1987, 88) and the analogous multivariate structural models of Harvey (1986, 1989) are a multivariate extension of the DLM which, by imposing rather stringent restrictions on the structure of the model, allows the observation covariance structure to be estimated on-line in a way analogous to the univariate analysis.

Suppose that θ_{tj} is the s -dimensional state vector for Y_{tj} , $j = 1, \dots, n$ then the DMR model specifies that for each j , Y_{tj} has a univariate DLM with the following observation and system equations:

Observation equation

$$Y_{ij} = F_t^T \theta_{ij} + v_{ij}, \quad v_{ij} \sim N(0, V_t \sigma_j^2), \quad (3.9)$$

System equation

$$\theta_{ij} = G_t \theta_{(t-1)j} + w_{ij}, \quad w_{ij} \sim N(0, W_t \sigma_j^2), \quad (3.10)$$

Initial information

$$\left. \begin{aligned} (\theta_{0j} | D_0, \sigma_j^{-2}) &\sim N(m_{0j}, \sigma_j^2 C_0) \\ (\sigma_j^{-2} | D_0) &\sim G\left(\frac{n_0}{2}, \frac{n_0 S_{0j}}{2}\right) \end{aligned} \right\} \quad (3.11)$$

All the defining quantities of the model are assumed known except for σ_j^2 which is some unknown variance scale factor. Each series typically has a different state vector θ_{ij} , state mean m_{0j} and expected variance S_{0j} for $j = 1, \dots, n$. On the other hand, F_t , G_t , V_t , W_t , C_0 and n_0 are assumed to be the same for each of the n series. The error sequences $\{w_{ij}\}_{i \geq 1}$ and $\{v_{ij}\}_{i \geq 1}$ are both independent through time and mutually independent of each other for all j .

Notice that when $V_t = 1$ the univariate DLM for Y_{ij} defined above is equivalent to the DLM defined in section 3.3 where the univariate observation variance is unknown. Also note that central to the theory of DMR models is the fact that each univariate distribution has *common* F_t and G_t . This makes the model appropriate for multivariate series which exhibit a high degree of symmetry in all but their location. Any common measurement scales can be accounted for by the common observation scale factor V_t .

Define Σ to be the unknown covariance matrix given by:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_n^2 \end{pmatrix}$$

so that for each i :

$$\text{var}(v_{ti} | \Sigma) = V_t \sigma_i^2, \quad \text{cov}(w_{ti} | \Sigma) = W_t \sigma_i^2,$$

and for all $i \neq j$:

$$\begin{aligned}\text{cov}(v_{ti}, v_{tj} | \Sigma) &= V_t \sigma_{ij} \\ \text{cov}(w_{ti}, w_{tj} | \Sigma) &= W_t \sigma_{ij}.\end{aligned}$$

Now the DMR model considers the state parameter vectors and evolution error vectors collectively in $(s \times n)$ matrices such that

$$\begin{aligned}\Theta_t &= (\theta_{t1}, \dots, \theta_{tn}) \\ \Omega_t &= (w_{t1}, \dots, w_{tn}).\end{aligned}$$

Let $v_t^T = (v_{t1}, \dots, v_{tn})$, then Y_t follows a DMR model if the following observation, system equations and prior distribution for Θ_t and Σ hold for all time:

Observation equation

$$Y_t^T = F_t^T \Theta_t + v_t^T, \quad v_t \sim N(\mathbf{o}, V_t \Sigma)$$

System equation

$$\Theta_t = G_t \Theta_{t-1} + \Omega_t, \quad \Omega_t \sim N(0, W_t, \Sigma)$$

Initial information

$$(\Theta_0, \Sigma | D_0) \sim NW_{n_0}^{-1}(M_0, C_0, n_0 S_0)$$

F_t , G_t , V_t , W_t , C_0 and n_0 are unchanged from equations 3.9, 3.10 and 3.11; $M_0 = (m_{01}, \dots, m_{0n})$ is an $s \times n$ matrix and S_0 is an $n \times n$ positive definite matrix with $\{S_{01}, \dots, S_{0n}\}$ of distributions 3.11 on its diagonal. Notice that the model uses a *row* vector of observations and a *matrix* of state parameters. This special arrangement utilises the fact that each series shares the same F_t and G_t to give a simple conjugate multivariate analysis.

Ω_t is said to have a matrix-variate normal distribution, (see, for example, Dawid, 1981), whose density is given by:

$$p\{\Omega_t\} = k\{W_t, \Sigma\} \exp \left[-\frac{1}{2} \text{trace} \{ \Omega_t^T W_t^{-1} \Omega_t \Sigma^{-1} \} \right]$$

where

$$k\{W_t, \Sigma\} = (2\pi)^{-\frac{ns}{2}} |W_t|^{-\frac{n}{2}} |\Sigma|^{-\frac{s}{2}}.$$

The joint posterior distribution of Θ_{t-1} and Σ at time $t-1$ (and hence the initial distribution at time 0) is said to be a normal/inverse Wishart distribution such that:

$$(\Theta_{t-1} | \Sigma, \mathbf{y}^{t-1}) \sim N(M_{t-1}, C_{t-1}, \Sigma), \quad (\Sigma | \mathbf{y}^{t-1}) \sim W_{n_{t-1}}^{-1}(n_{t-1} S_{t-1}) \quad (3.12)$$

for some M_{t-1} , C_{t-1} , n_{t-1} and S_{t-1} . Following Dawid (1981), Σ is said to have an inverse Wishart distribution if the density for Σ is given by:

$$p(\Sigma | \mathbf{y}^{t-1}) = c |\Sigma|^{-(n+s/2)} \exp \left[-\frac{1}{2} \text{trace}(n_{t-1} S_{t-1} \Sigma^{-1}) \right]$$

where c is a normalising constant and $E(\Sigma | \mathbf{y}^{t-1}) = (S_{t-1} n_{t-1}) / (n_{t-1} - 2)$ for $n > 2$.

The marginal posterior distribution of Θ_{t-1} at time $t-1$ is known as a matrix T distribution (Dawid, 1981) and is denoted by:

$$(\Theta_{t-1} | \mathbf{y}^{t-1}) \sim T_{n_{t-1}}(M_{t-1}, C_{t-1}, n_{t-1} S_{t-1}).$$

As for the matrix normal distribution each component of Θ_{t-1} has a multivariate T distribution with n_{t-1} degrees of freedom so that

$$(\theta_{(t-1)j} | \mathbf{y}^{t-1}) \sim T_{n_{t-1}}(m_{(t-1)j}, C_{t-1} S_{(t-1)j}).$$

The updating equations and forecast distributions for the DMR model are given as follows (for full details see Quintana, 1985, 87). The joint posterior

distribution of Θ_{t-1} and Σ given in equation 3.12 gives the joint prior distribution at time t :

$$(\Theta_t, \Sigma | y^{t-1}) \sim NW_{n_{t-1}}^{-1}(a_t, R_t, n_{t-1}S_{t-1})$$

where

$$a_t = G_t M_{t-1} \quad \text{and} \quad R_t = G_t C_{t-1} G_t^T + W_t.$$

This leads directly to the one-step ahead forecast distribution for Y_t conditional on Σ :

$$(Y_t | \Sigma, y^{t-1}) \sim N(f_t, Q_t \Sigma)$$

with margin:

$$(Y_t | y^{t-1}) \sim T_{n_{t-1}}(f_t, Q_t S_{t-1}) \quad (3.13)$$

where

$$f_t^T = F_t^T a_t \quad \text{and} \quad Q_t = V_t + F_t^T R_t F_t.$$

After observing y_t the joint distribution of Θ_t and Σ can be updated to give the posterior distribution:

$$(\Theta_t, \Sigma | y^t) \sim NW_{n_t}^{-1}(M_t, C_t, n_t S_t)$$

where

$$\begin{aligned} M_t &= a_t + A_t e_t^T & \text{and} & \quad C_t = R_t - A_t A_t^T Q_t, \\ n_t &= n_{t-1} + 1 & \text{and} & \quad S_t = n_t^{-1} (n_{t-1} S_{t-1} + e_t e_t^T / Q_t) \end{aligned}$$

where

$$A_t = R_t F_t / Q_t \quad \text{and} \quad e_t = y_t - f_t.$$

Notice that a_t and M_t are both $(s \times n)$ matrices and their j^{th} columns contain the means of the state parameters of the DLM of Y_{tj} . Also note that as F_t , G_t , V_t , W_t and C_0 are common by definition of the DMR model, then the $(s \times s)$ matrices R_t and C_t , the $(s \times 1)$ adaptive coefficient A_t and the scalar Q_t are common to the n series also.

One property which emphasises the weakness of the dependence structure across the components in the series, is that the one step-ahead forecast distribution of a component $Y_t(i)$ depends only on the history of that component and not on Y^{t-1} (West & Harrison, 1989a).

Note that if $n = 1$ then the DMR model reduces to a DLM with unknown variance. If Σ is diagonal, then the multivariate model is simply n unrelated univariate ones. Indeed, even when Σ is not diagonal, the only difference between the DMR model and n univariate DLM's with unknown variances, is that the covariance structure of the series is also estimated on-line.

Chapter 4

An Introduction to Graphical Models.

A multivariate random variable may have a set of conditional independence relationships across its component variables. Graphical models accommodate these conditional independences and represent them by a graphical structure.

This chapter gives a short review of some of the graph theoretic results which are used later in the thesis.

4.1 Influence Diagrams.

Let a general set of random vectors be denoted by X , Y , Z and W . The following conditional independence (c.i.) properties can then be defined on them:

$$\text{P1)} \quad X \perp\!\!\!\perp Y \mid (Y, Z)$$

$$\text{P2)} \quad X \perp\!\!\!\perp Y \mid Z \Leftrightarrow Y \perp\!\!\!\perp X \mid Z$$

$$\text{P3)} \quad X \perp\!\!\!\perp (Y, Z) \mid W \Leftrightarrow \begin{cases} X \perp\!\!\!\perp Y \mid (Z, W) \\ \text{together with} \\ X \perp\!\!\!\perp Z \mid W \end{cases}$$

where $X \perp\!\!\!\perp Y \mid Z$ reads “ X is independent of Y given Z ” (see Dawid, 1979).

Let $X = (X_1, \dots, X_m)$ be an ordered set of random vectors on which $\cdot \perp\!\!\!\perp \cdot \mid \cdot$ is defined. Suppose the following $m - 1$ conditional independence statements are given:

$$X_r \perp\!\!\!\perp Q(X_r) \mid P(X_r) \quad 2 \leq r \leq m \quad (4.1)$$

where $P(X_r) \subseteq \{X_1, \dots, X_{r-1}\}$ and $Q(X_r) = \{X_1, \dots, X_{r-1}\} \setminus P(X_r)$ are vectors of components from (X_1, \dots, X_{r-1}) , such that for two sets A and B , $A \setminus B$ denotes those elements in A which are *not* in B .

The motivation for considering sets of such c.i. statements is as follows. Suppose that X_1, \dots, X_m have a joint density $p(\mathbf{x}_1, \dots, \mathbf{x}_m)$. Then this can be written in the product form such that:

$$p(\mathbf{x}_1, \dots, \mathbf{x}_m) = \prod_{i=1}^m p(\mathbf{x}_i | \mathbf{x}_1, \dots, \mathbf{x}_{i-1}).$$

Using equation 4.1 it can be seen that $\mathbf{x}_1, \dots, \mathbf{x}_{i-1}$ can be replaced by $P(\mathbf{x}_i)$ without loss. Thus the joint density can be simplified and rewritten as:

$$p(\mathbf{x}) = \prod_i p(\mathbf{x}_i | P(\mathbf{x}_i)).$$

A set of conditional independence statements such as these can be usefully represented graphically. A graph $G(X, E)$ comprises a set of *nodes* X and *edges* E , where a directed edge from X_i to X_j is denoted by $(X_i, X_j) \in E$. A directed graph with a directed edge from X_i to X_r iff $X_i \in P(X_r)$, is called the *graph of an influence diagram* – see Howard and Matheson (1981), Shachter (1986) and Smith (1989, 90). The set $P(X_r)$ is known as the *parent set* of X_r . The graph of an influence diagram, together with the c.i. statements 4.1 is called an *influence diagram* (ID).

For example, suppose there are 4 variables, X_1 , X_2 , X_3 and X_4 with the following conditional independence statements defined across them:

$$X_3 \perp\!\!\!\perp X_1 | X_2$$

$$X_4 \perp\!\!\!\perp X_1 | X_2, X_3.$$

These conditional independences can be represented by the graph of an influence

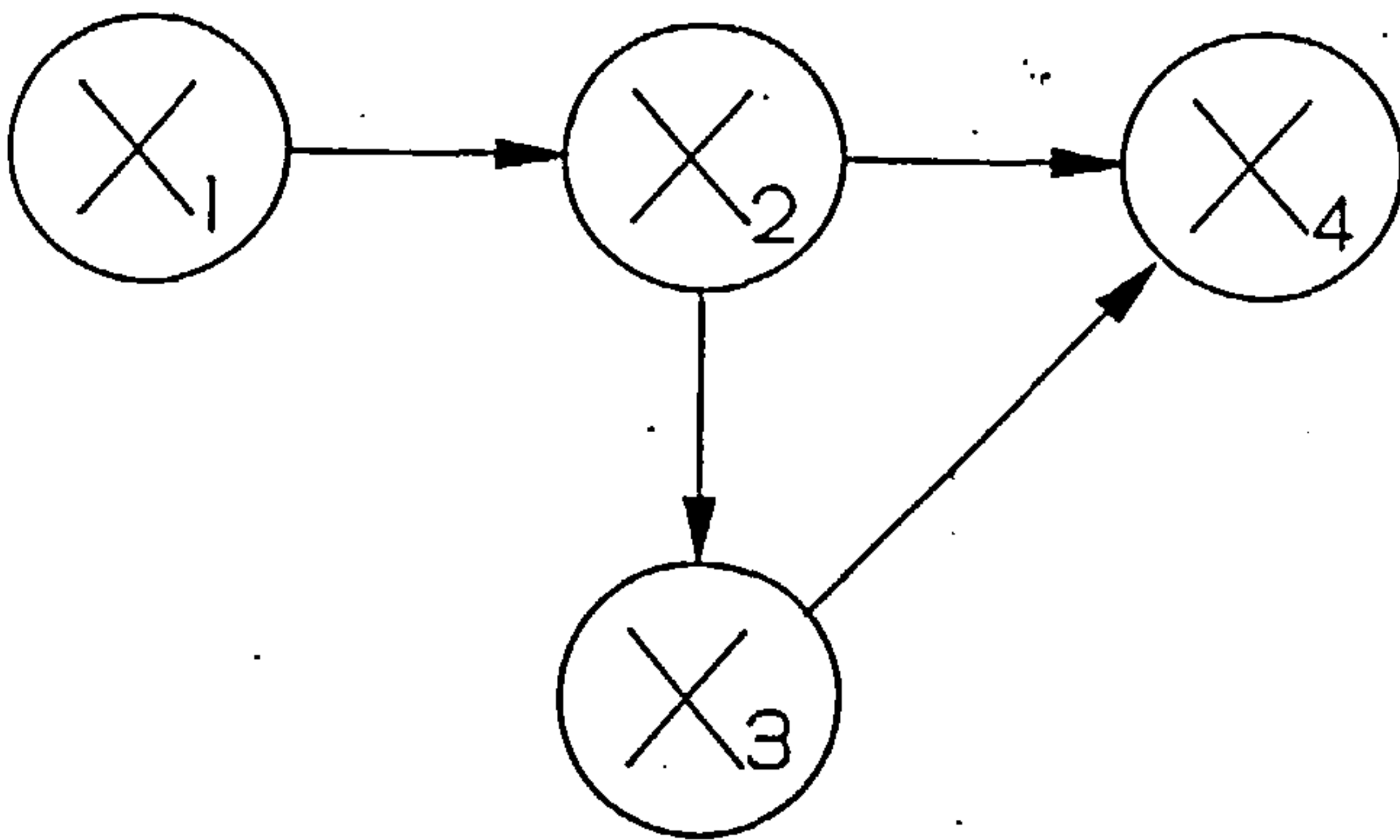


Figure 4.1: The graph of an influence diagram defined on 4 variables.

diagram shown in figure 4.1. So

$$\begin{array}{ll}
 P(X_2) = \{X_1\} & Q(X_2) = \emptyset \\
 P(X_3) = \{X_2\} & Q(X_3) = \{X_1\} \\
 P(X_4) = \{X_2, X_3\} & Q(X_4) = \{X_1\}
 \end{array}$$

A *path* of length m from X_i to X_k is a sequence of nodes, $X_i = X_{i,0}, \dots, X_{i,m} = X_k$, where $(X_{i,(j-1)}, X_{i,j}) \in E$ for each $j = 1, \dots, m$. The set of vertices with paths to X_k , together with X_k , is called the *ancestor set* of X_k and is denoted by $an(X_k)$. For example, $an(X_3)$ in figure 4.1 is $\{X_1, X_2, X_3\}$.

Although in this thesis conditional independence will be of the usual type, many of the results proved here extend to generalised conditional independence structures (Smith, 1989), that is, sets of objects on which a ternary operator $\cdot \amalg \cdot | \cdot$ can be defined which satisfy properties $P1$, $P2$ and $P3$. A simple example of this is when X , Y and Z are random vectors but $X \amalg Y | Z$ reads “a best linear estimate (under quadratic loss) of the components of X based on the components of Y and Z need only include the components of Z ” (for a proof see Smith (1989)).

It is straightforward to check (Smith, 1989) that if the random vectors X_1, \dots, X_m have their order permuted to $X_{i(1)}, \dots, X_{i(m)}$ in a way which is compatible with the graph of the influence diagram G of X_1, \dots, X_m , then the conditional inde-

pendence statements associated with G and X_1, \dots, X_m are the same as with $X_{i(1)}, \dots, X_{i(m)}$. In this case the graph of the influence diagram *without* numbering of variables, contains sufficient information to retrieve all input c.i. statements.

4.2 Directed Markov separation.

The influence diagram gives a set of $m - 1$ “local” (or “pairwise”) Markov statements. Further conditional independence statements can be deduced from its graph using properties P1, P2 and P3. These are sometimes called the *global Markov* properties of the system.

The following theorem was proved in a rather obscure way by Pearl & Verma (1987), Pearl (1988) and later simplified into the given form by Lauritzen et al (1990). It identifies all the conditional independences defined by an influence diagram and gives an algorithm for deciding whether any given conditional independence can be logically deduced from an influence diagram.

Theorem 4.2.1 *Suppose that conditional independence statements for a set of variables are represented in an influence diagram whose graph is I and that U , V and W denote sets of variables on it. Adapt the influence diagram in the following way:*

- *Form the directed subgraph I_1 of I whose nodes are in the ancestor set $an(U, V, W)$ and whose directed edges are those in I which lie between these nodes.*
- *Join all pairs of nodes $(X, Y) \in P(Z)$ by an undirected edge, where $P(Z)$ is the parent set of Z and $Z \in I_1$. This process is known as moralising the*

graph since all parents of a single variable are joined by an arc. Call this mixed graph I_2 .

- Form an undirected graph J by replacing all directed edges in I_2 by undirected edges.

Then

$$U \amalg V | W$$

if all undirected paths in J between a node $R \in U$ and $S \in V$, must pass through a node $T \in W$.

Furthermore, if this condition is violated, then there exists a probabilistic influence diagram respecting the c.i. statements of the influence diagram, for which $U \amalg V | W$ is not true.

Consider the graph of the influence diagram I in figure 4.2. Suppose that we would like to know whether $U \amalg V | W$. First of all the graph I_1 given in b) of figure 4.2 is derived. Notice how nodes c, d, e, f and g have been deleted as none of them have paths leading to U, V or W . Graph J in figure 4.2 gives the final undirected moralised graph of I . It is now simple to see that it is not true that:

$$U \amalg V | W$$

as there are two paths (U, a, b, V) and (U, b, V) which do not pass through W .

An interesting question to ask is when do two influence diagrams on the same set of random vectors (X_1, \dots, X_m) imply the same set of c.i. statements. Verma & Pearl (1990) report the following result as simply deducible from the theorem 4.2.1.

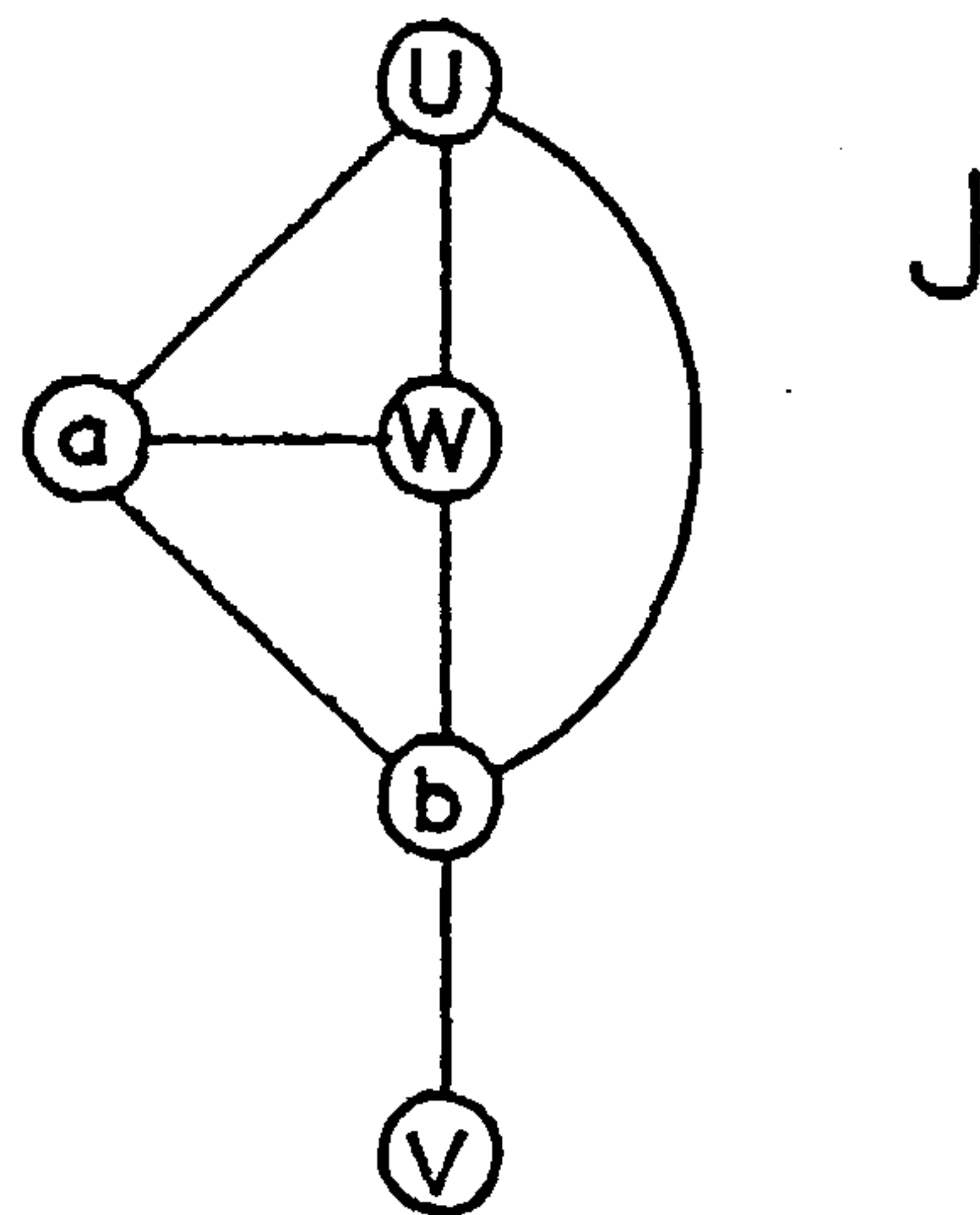
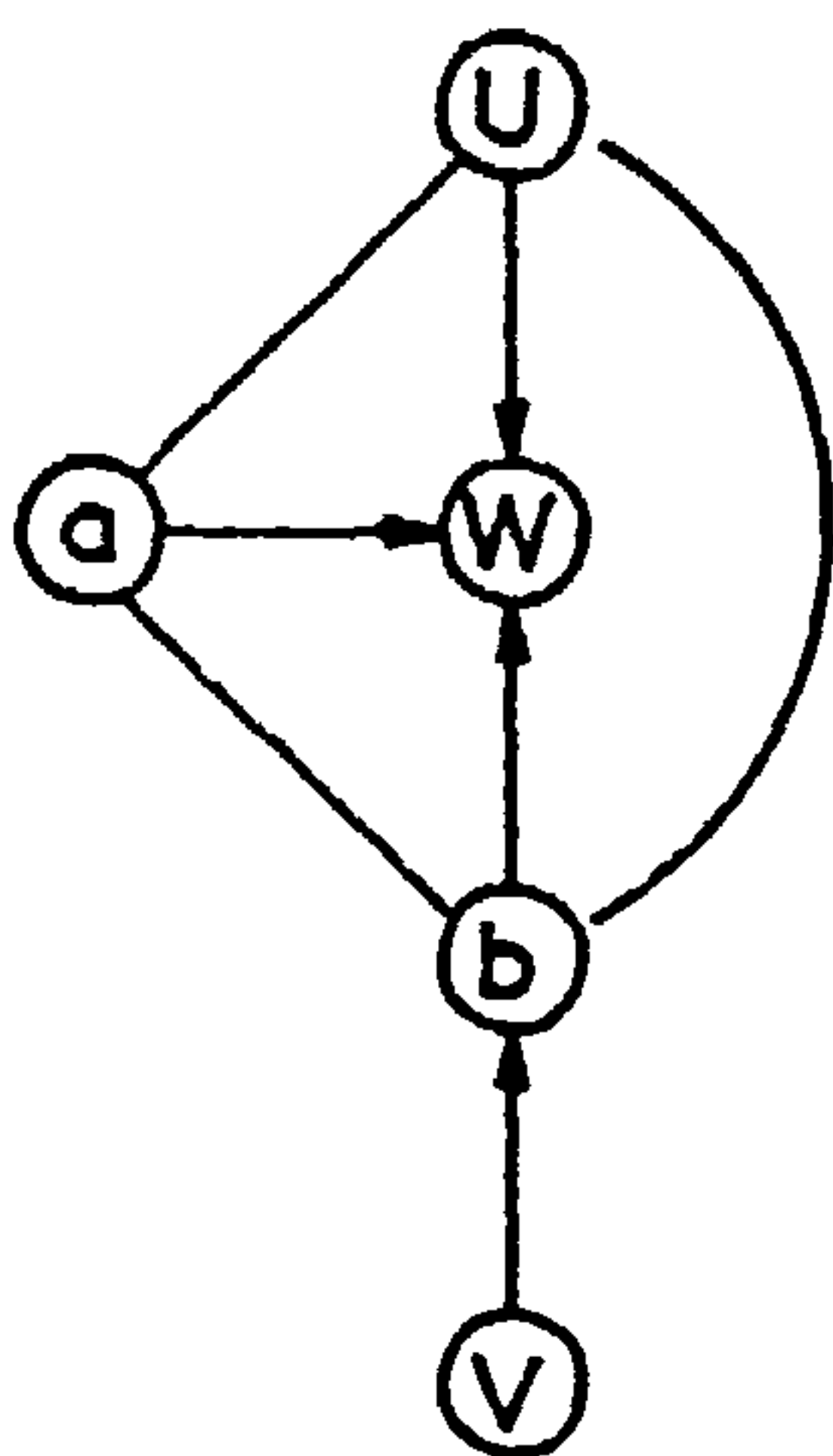
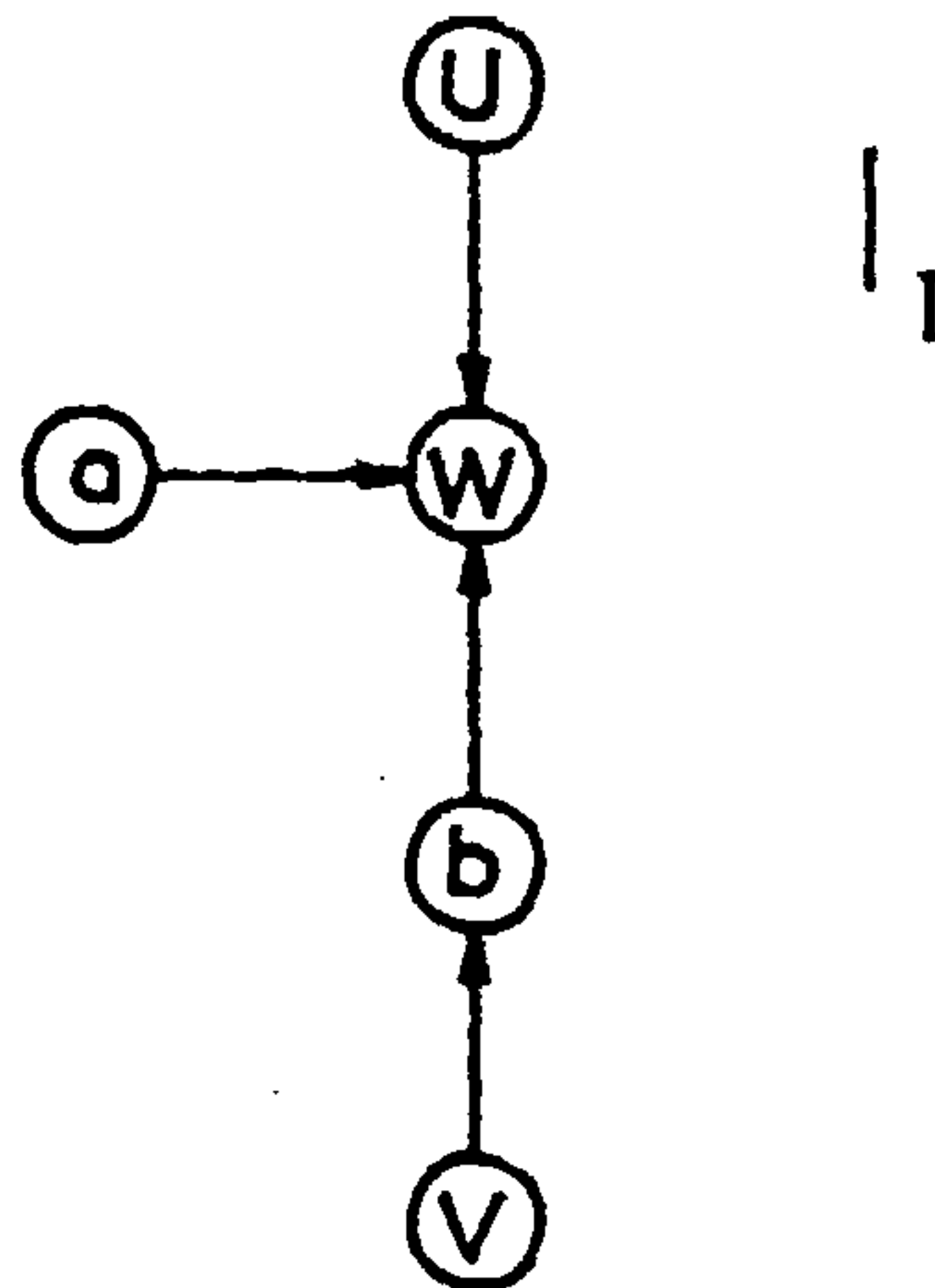
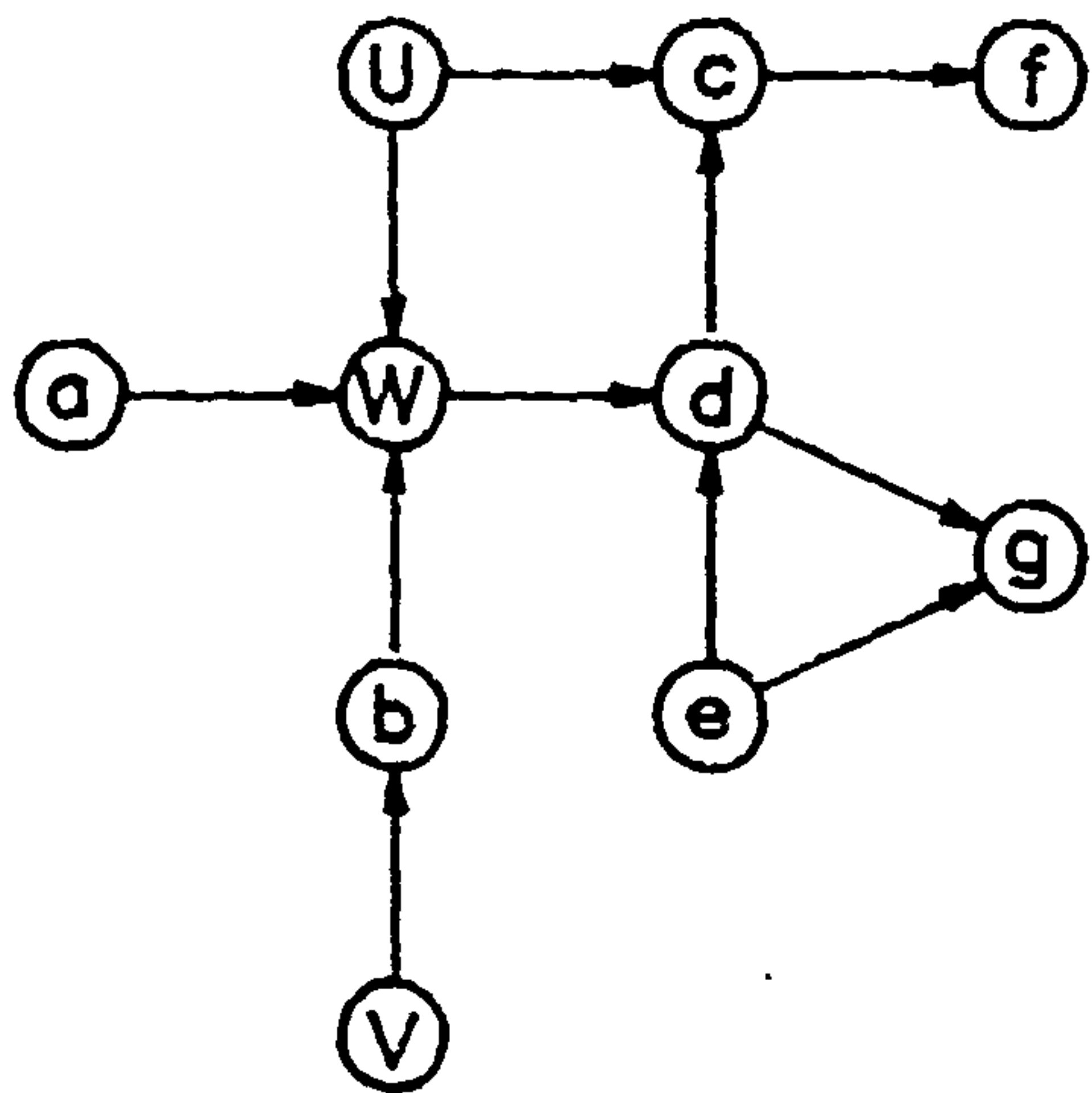


Figure 4.2: Illustration of theorem 4.2.1. Is $U \perp\!\!\!\perp V \mid W$?

Theorem 4.2.2 *Two influence diagrams are equivalent, that is, they have the same implied conditional independence structure, iff:*

1. *their undirected graphs are the same*

and

2. *both graphs share the same unmarried parents.*

Thus it is straightforward to decide whether two influence diagrams on the same set of variables share the same c.i. statements.

4.3 Decomposable Graphs.

Decomposable graphs are influence diagrams of a special form. They have several properties which make them a particularly useful class of ID to study. Firstly they allow simple arc reversal which in turn makes it very easy to determine other ID's having the same structure (Smith, 1989). Secondly the conditional independence relationships on the undirected version of a decomposable graph can be found directly (Smith, 1989, Kiiveri, Speed & Carlin, 1984). These results follow directly from theorem 4.2.1. Finally, they make it particularly simple to propagate probabilities as the joint distribution can be stored as margins on cliques (Lauritzen & Spiegelhalter, 1988). A decomposable graph will now be formally defined. There are various definitions of decomposable graphs by different authors, each author concentrates on a particular equivalent property.

Call a graph of an influence diagram *decomposable* if all parents in its graph are *married* (that is, all parents are joined by an edge). Figure 4.3 contains 2 graphs, graph *G* is decomposable but graph *H* is not, since there is no edge between the parent nodes *a* and *b*.

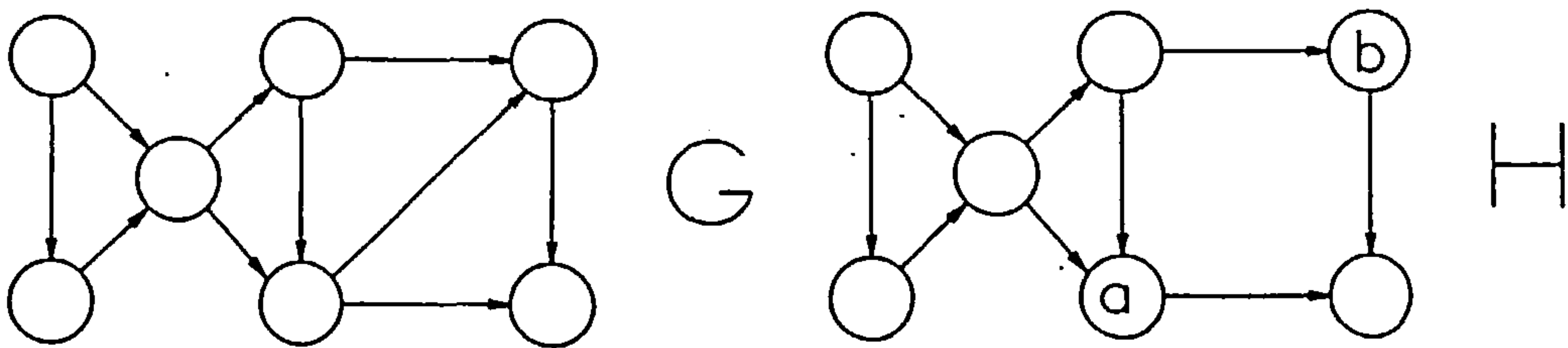


Figure 4.3: Graph of the influence diagram G is decomposable, whereas H is not because parent nodes a and b are not joined.

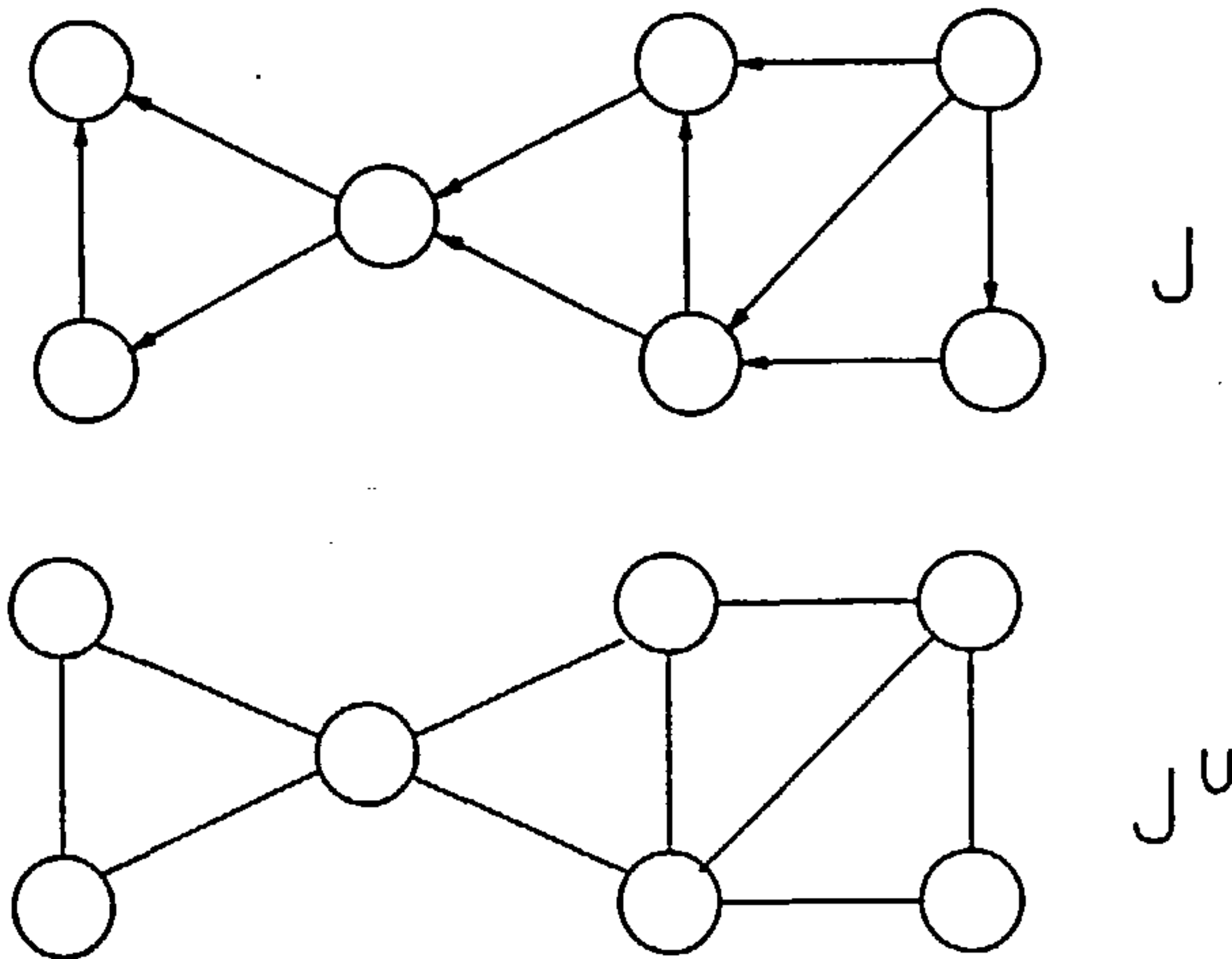


Figure 4.4: Decomposable graph J together with J^u .

An *undirected* graph is defined by a set of nodes together with an edge set E which has the property that if $(X_i, X_j) \in E$ then so is (X_j, X_i) . If G is the graph of a decomposable influence diagram, then let G^u be the undirected graph formed from G where all directed edges are replaced by undirected ones. Smith (1989) calls two decomposable influence diagrams G_1, G_2 *similar* if $G_1^u = G_2^u$. The graph of a decomposable ID J together with J^u are given in figure 4.4. Notice how J^u and G^u of figure 4.3 are identical, thus J and G are similar decomposable ID's.

An undirected graph is *triangulated* if for any cycle of nodes $(X_{i(0)}, X_{i(1)}, \dots, X_{i(k)} = X_{i(0)})$, $k \geq 4$, there is a *chord*, that is, there is some edge $(X_{i(j)}, X_{i(l)})$ where $j \neq l \pm 1 \pmod{k+1}$. Triangulated graphs have been described by Berge (1973), Golumbic (1980), Dirac (1961) under the name of rigid circuits,

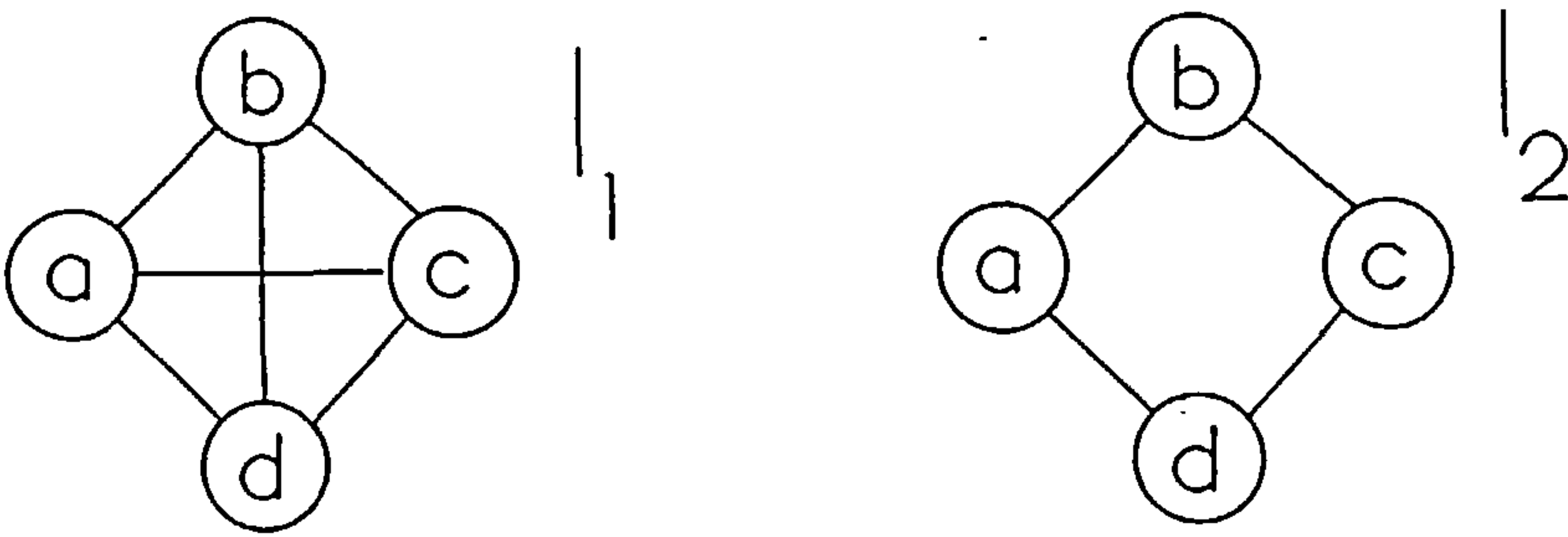


Figure 4.5: Two graphs such that I_1 is triangulated and I_2 is not.

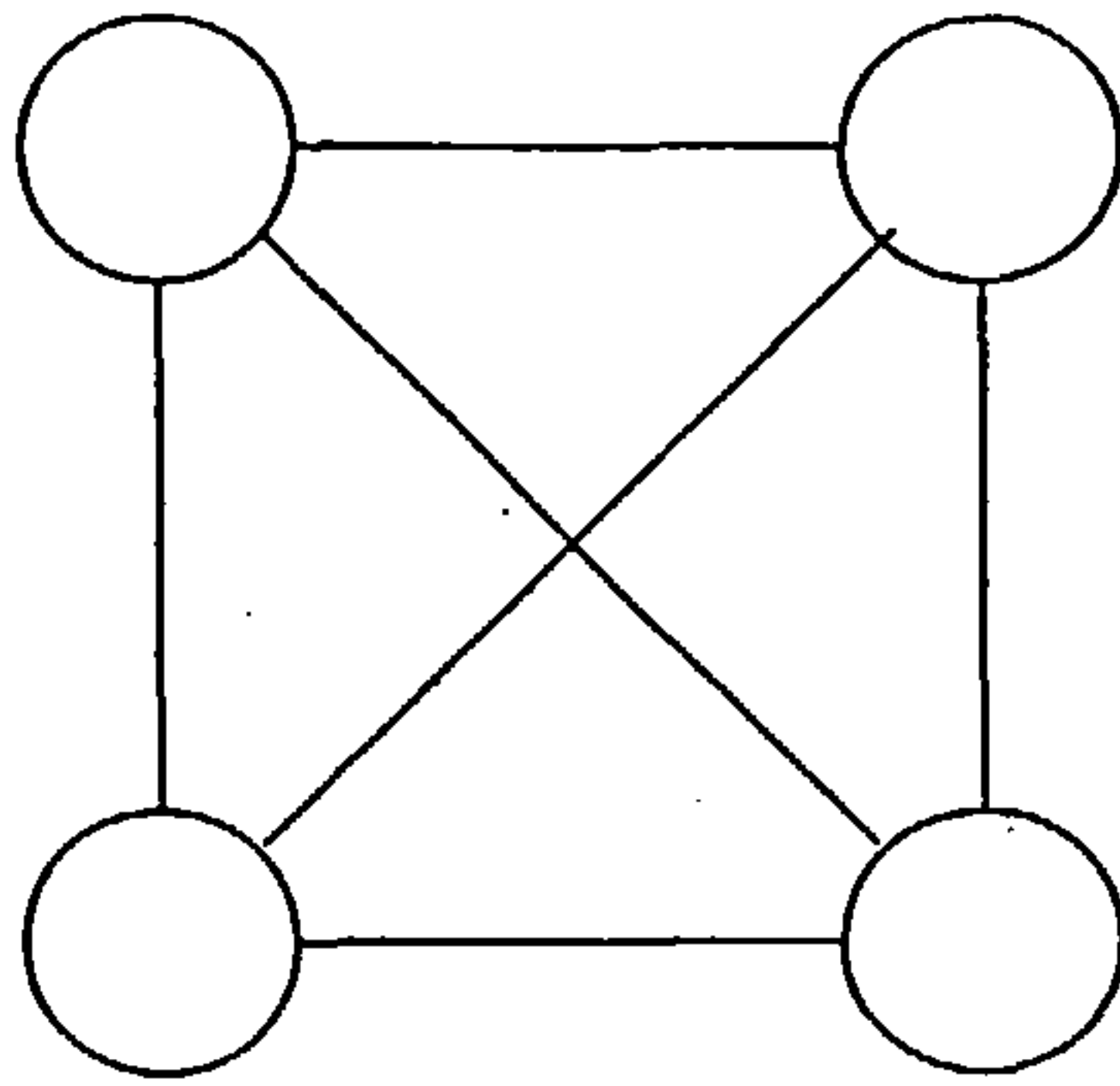


Figure 4.6: A complete graph.

Gavril (1972) who calls them chordal, and Lauritzen et al (1984). The graphs in figure 4.5 both have the cycle (a, b, c, d) . Graph I_1 is triangulated with chords $\{a, c\}$ and $\{b, d\}$, whereas graph I_2 is not triangulated since it has no chords.

A *complete graph* is one in which there is an edge between every pair of nodes in a graph. Figure 4.6 shows an example of a complete graph. A *clique* is a maximally connected subgraph \overline{G} of G , that is, \overline{G} is complete and is not contained in any other complete subgraph of G . For example, the graph in figure 4.7 has the cliques $\{X_1, X_2, X_3\}$, $\{X_2, X_3, X_4\}$, $\{X_3, X_4, X_5\}$ and $\{X_5, X_6\}$.

An undirected graph is said to have the *running intersection property* (RIP) (Beeri, 1981, 83, Lauritzen et al, 1984, Tarjan & Yannakakis, 1984) if the cliques

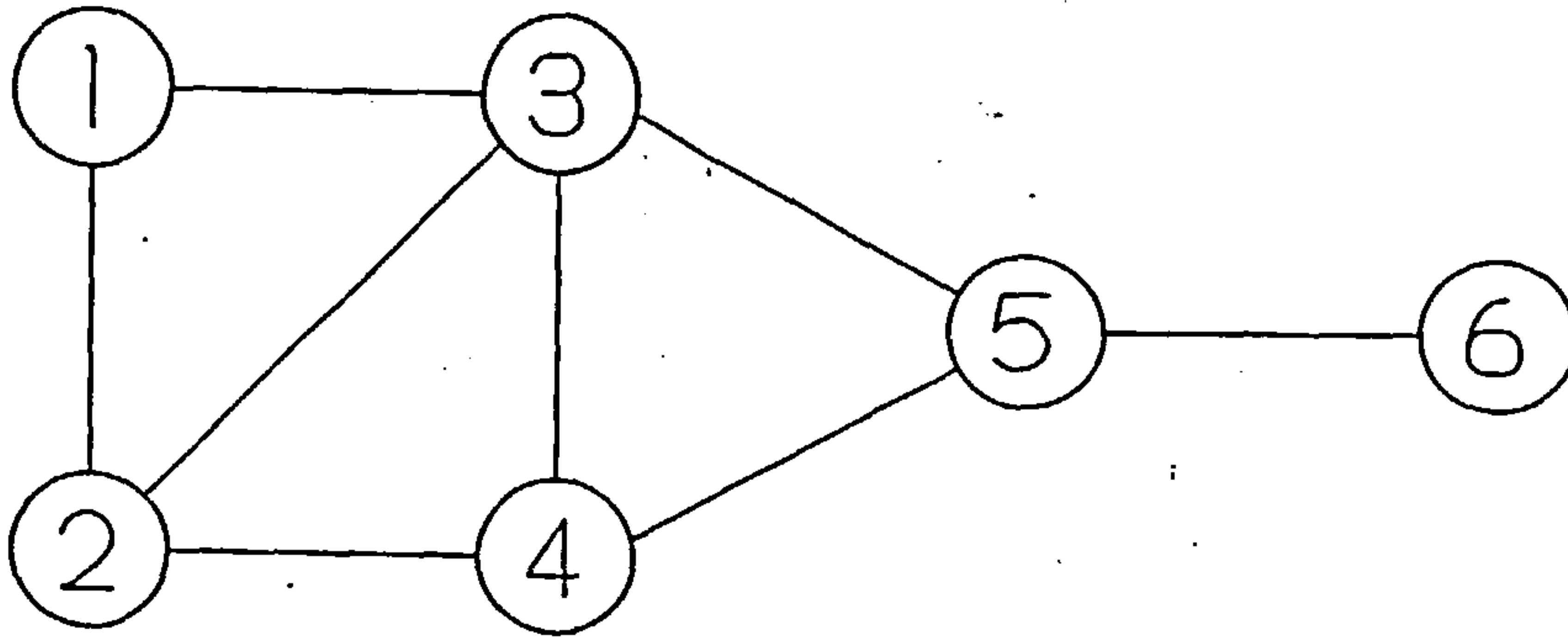


Figure 4.7: Graph with the cliques $\{1,2,3\}$, $\{2,3,4\}$, $\{3,4,5\}$ and $\{5,6\}$.

on the graph can be indexed $C(1), \dots, C(n)$ so that

$$S(l) = \left(C(l) \cap \bigcup_{i=1}^{l-1} C(i) \right) \subseteq C(p(l))$$

for some $p(l)$, $1 \leq p(l) \leq l-1$ this being true for $l = 2, \dots, n$. That is, the intersection of the l^{th} clique with all the preceding ones is contained in just one of the preceding cliques. The undirected graph in figure 4.8 has the RIP on the cliques $C(1) = \{X_1, X_2, X_3, X_4\}$, $C(2) = \{X_3, X_4, X_5\}$, $C(3) = \{X_5, X_6\}$ and $C(4) = \{X_3, X_5, X_7\}$, since:

$$S(2) = C(2) \cap C(1) = \{X_3, X_4\} \subseteq C(1)$$

$$S(3) = C(3) \cap (C(1) \cup C(2)) = \{X_5\} \subseteq C(2)$$

$$S(4) = C(4) \cap (C(1) \cup C(2) \cup C(3)) = \{X_3, X_5\} \subseteq C(2)$$

Note that in this example $p(2) = 1$ and $p(3) = p(4) = 2$.

The following theorems, the first of which is given by Smith (1989) are a direct consequence of applying theorem 4.2.1.

Theorem 4.3.1 *If two decomposable graphs are similar, then they contain equivalent conditional independence statements, i.e., all conditional independence statements deducible from one are deducible from the other.*

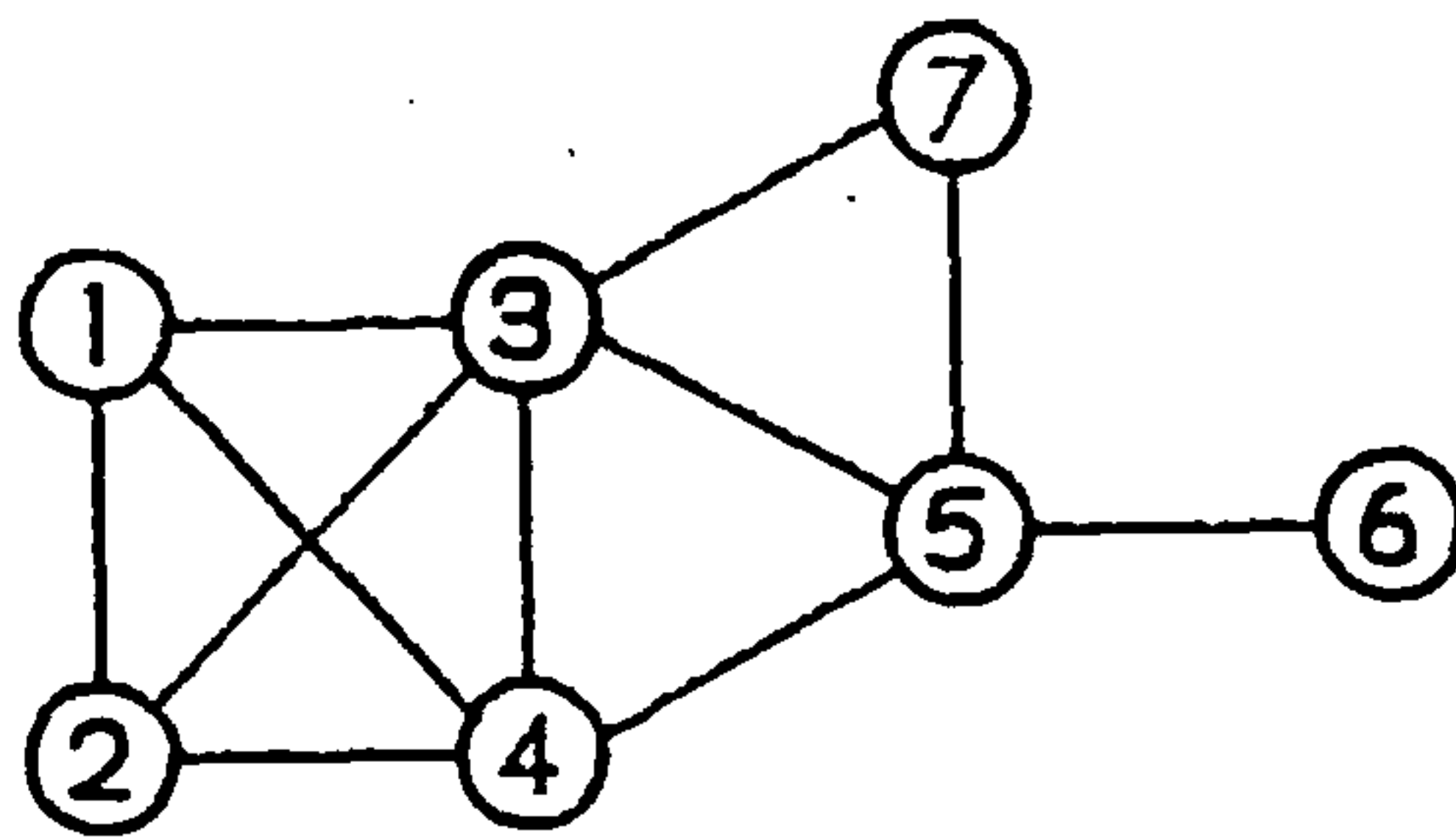


Figure 4.8: An undirected graph following the RIP.

Theorem 4.3.2 *The following statements are equivalent:*

1. *The graph of the influence diagram G is decomposable.*
2. *The undirected graph G^u of G is triangulated.*
3. *There is an ordering of cliques $(C(1), \dots, C(n))$ such that G^u has the RIP.*

Note that the ordering of the RIP for decomposable graphs is never unique. One way to construct the ordering of the cliques is to introduce the cliques in an order compatible with the ordering of the nodes in the ID. The choice of the first clique can be arbitrary and so there are at least n possible choices of ordering of cliques for a decomposable ID with graph G which exhibit the RIP. For example, once again consider the graph in figure 4.8. Now instead of having the cliques $C(1), \dots, C(4)$ given above, the cliques could have been ordered such that $C(1) = \{X_3, X_4, X_5\}$, $C(2) = \{X_3, X_5, X_7\}$, $C(3) = \{X_5, X_6\}$ and $C(4) = \{X_1, X_2, X_3, X_4\}$. Then $S(2) = \{X_3, X_5\} \subseteq C(1)$, $S(3) = \{X_5\} \subseteq C(1)$ and $S(4) = \{X_3, X_4\} \subseteq C(1)$, and so with this new ordering of cliques $p(l) = 1$ for $l = 2, 3, 4$. Finally note that by the definition of equivalent graphs given in theorem 4.2.2, any two decomposable graphs which have the same undirected graph are equivalent.

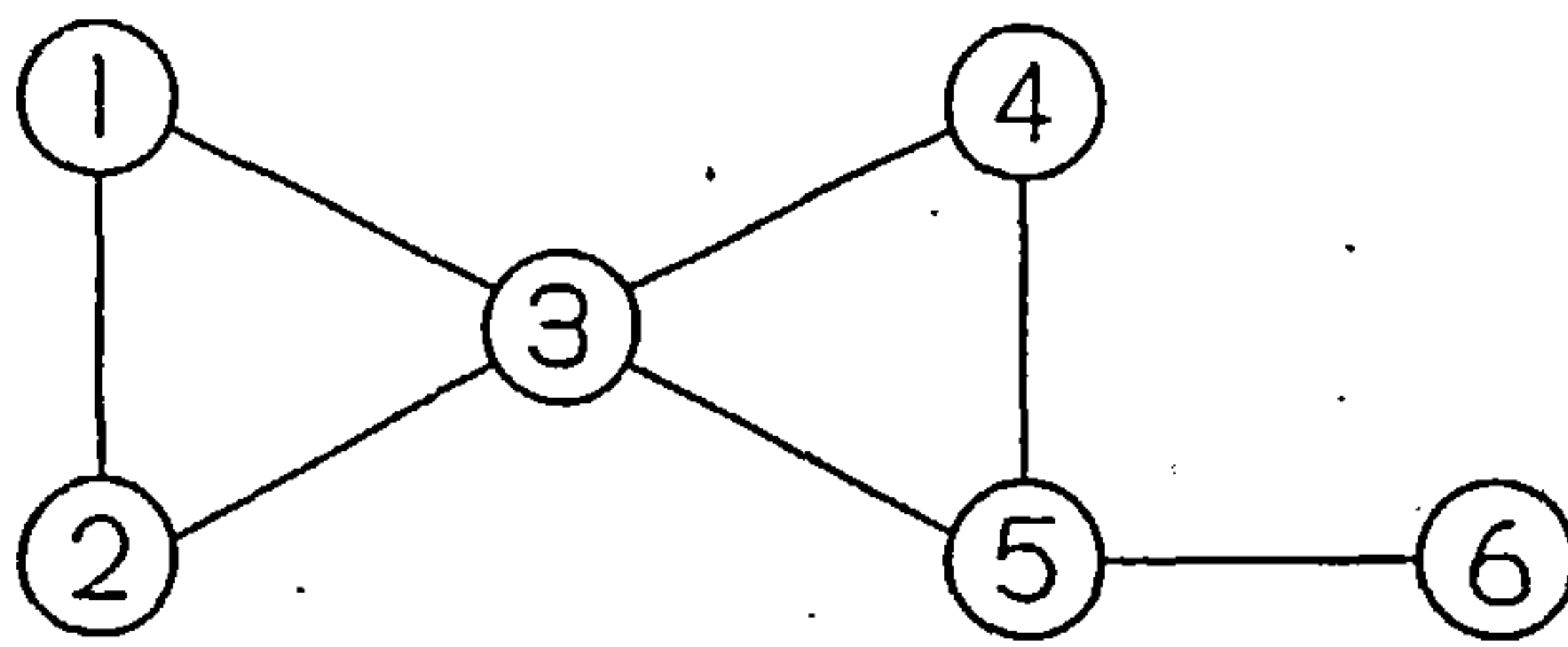


Figure 4.9: An undirected graph.

In an undirected graph, if $(X_i, X_j) \in E$ then X_j is said to be a *neighbour* of X_i . The set of all neighbours of X_i is denoted $n(X_i)$. The conditional independence statements in an undirected graph are interpreted as follows:

$$X_i \perp\!\!\!\perp X \setminus \{X_i, n(X_i)\} \mid n(X_i).$$

For example, the conditional independence statements of the graph in figure 4.9 are as follows:

$$1 \perp\!\!\!\perp \{4, 5, 6\} \mid \{2, 3\}$$

$$2 \perp\!\!\!\perp \{4, 5, 6\} \mid \{1, 3\}$$

$$3 \perp\!\!\!\perp \{6\} \mid \{1, 2, 4, 5\}$$

$$4 \perp\!\!\!\perp \{1, 2, 6\} \mid \{3, 5\}$$

$$5 \perp\!\!\!\perp \{1, 2\} \mid \{3, 4, 6\}$$

$$6 \perp\!\!\!\perp \{1, 2, 3, 4\} \mid \{5\}$$

Smith (1989) introduced a result which states that if a decomposable influence diagram I has ordered nodes (X_1, \dots, X_m) , then

$$X_i \perp\!\!\!\perp X \setminus \{X_i, n(X_i)\} \mid n(X_i).$$

This means that given a directed decomposable ID, the c.i. relationships associated with an undirected graph G^u can be directly deduced from any influence diagram whose undirected graph is also G^u .

There is another interesting point to note. Suppose that the additional property:

$$\text{P4) } \left. \begin{array}{l} X \perp\!\!\!\perp Y | W, Z \\ \text{and} \\ X \perp\!\!\!\perp W | Y, Z \end{array} \right\} \Rightarrow X \perp\!\!\!\perp (W, Y) | Z.$$

is assumed on the variables. Then the conditional independences defined by an undirected graph G^u imply those of any decomposable influence diagram whose undirected graph is G^u . This will in particular be true if (X_1, \dots, X_m) are absolutely continuous and their joint density is strictly positive on the range of (X_1, \dots, X_m) .

Thus, not only can a decomposable graph give information about the c.i. statements on the corresponding undirected graph, but with the additional property P4, an undirected graph carries information about the c.i. structure in the corresponding decomposable influence diagrams.

4.4 Chain Graphs.

Chain graphs were introduced by Lauritzen & Wermuth (1984, 1989). In a *chain graph* the variables are partitioned into T subsets $\{V(1), \dots, V(T)\}$ so that

- $\{V(1), \dots, V(T)\}$ are ordered in a horizontal row,
- directed edges between subsets all go in the same direction
- and
- variables within $V(t)$, $t = 1, \dots, T$ can only be connected by undirected edges.

The set of *concurrent variables* is defined by $C(t) = V(1) \cup \dots \cup V(t)$. For any pair of unconnected vertices $\{x, y\}$ the *pairwise chain* Markov property given by

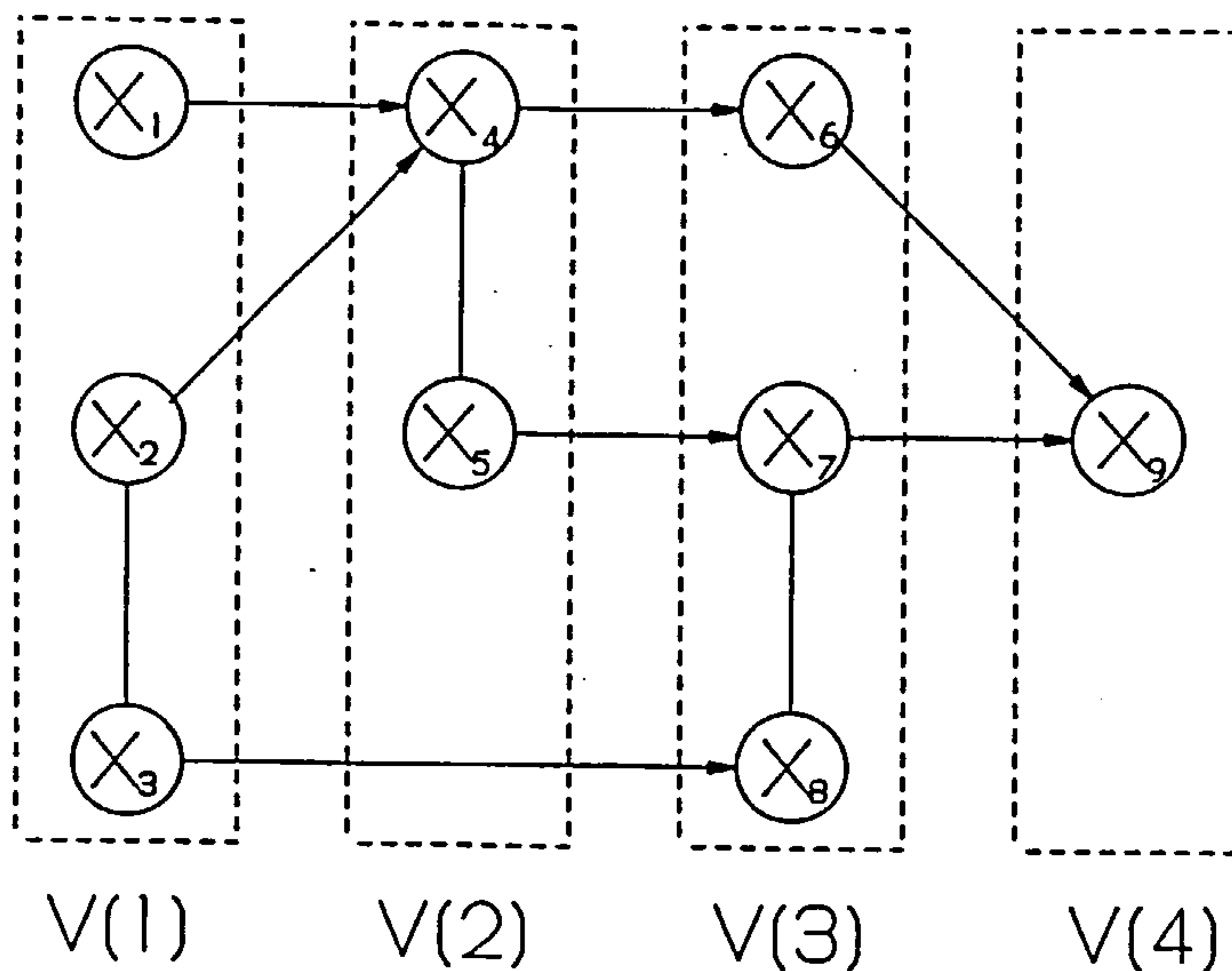


Figure 4.10: A chain graph with 4 subsets of variables $V(1)$, $V(2)$, $V(3)$ and $V(4)$.

Frydenberg (1989) states that:

$$x \perp\!\!\!\perp y \mid C(t^*) \setminus \{x, y\}$$

where t^* is the smallest t so that $\{x, y\} \in C(t)$, i.e., 2 unconnected variables are independent conditional on the set of concurrent variables containing them both. The ordered subsets are said to form a *dependence chain* on the subsets of concurrent variables. Notice that ID's are special cases of chain graphs where subsets contain just one element. Undirected graphical models are also a special case — this time the chain graph has just one subset containing all the variables.

Consider the chain graph in figure 4.10. Here $V(1) = \{X_1, X_2, X_3\}$, $V(2) = \{X_4, X_5\}$, $V(3) = \{X_6, X_7, X_8\}$ and $V(4) = \{X_9\}$. There are 4 sets of concurrent variables $V(1)$, $V(1) \cup V(2)$, $V(1) \cup V(2) \cup V(3)$ and $V(1) \cup V(2) \cup V(3) \cup V(4)$. Some of the pairwise chain Markov properties for this chain graph are as follows:

$$X_1 \perp\!\!\!\perp X_2 \mid X_3$$

$$X_6 \perp\!\!\!\perp X_7 \mid \{X_1, X_2, X_3, X_4, X_5, X_8\}$$

$$X_4 \perp\!\!\!\perp X_8 \mid \{X_1, X_2, X_3, X_5, X_6, X_7\}$$

Chain graph models are a family of distributions which are a mixture of both discrete and continuous variables with a set of conditional independence restrictions. Although they are very interesting in their own right, these models are not general enough for the purposes of modelling multivariate business time series of the type studied in this thesis. This is because each continuous variable is assumed to be Gaussian, conditional on its parents, and the dependence on the parent set only comes through the mean of the normal distribution. This implies joint normality across (X_1, \dots, X_m) which has been shown to be inappropriate for this application. However, later in the thesis, different classes of joint distribution are introduced which are also consistent with conditional independences represented through a chain graph and are more appropriate for these applications.

As with influence diagrams chain graphs can also be moralised. Here all pairs of nodes (X, Y) are joined by an undirected edge, if both X and Y are parents of nodes in the *same* subset of the chain graph.

It has been shown by Frydenberg (1989) that under the positivity condition P4), the global Markov properties of a chain graph can be derived so that for 3 sets of variables U , V and W :

$$U \perp\!\!\!\perp V \mid W$$

whenever W separates U and V in the moralised graph of $an(U, V, W)$, the smallest ancestor set containing the sets U , V and W .

For example, consider the chain graph and its corresponding moralised graph in a) and b) of figure 4.11. Edge $(2, 3)$ is added in the moralised graph because 2 and 3 are both parents of 4; edge $(1, 4)$ is added because 1 and 4 are parents

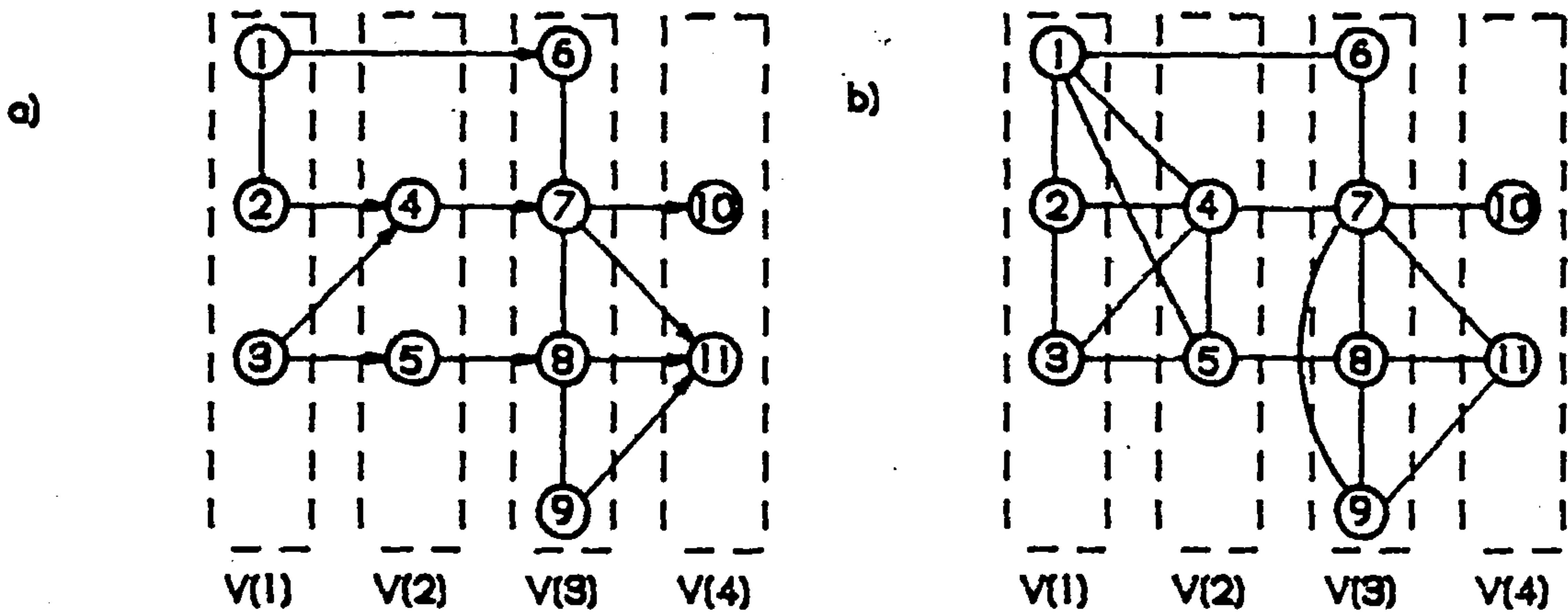


Figure 4.11: a) A chain graph and b) its moralised graph.

of 6 and 7 respectively where $\{6, 7\} \in V(3)$; edge $(1, 5)$ is added since 1 and 5 are parents of 6 and 8 respectively where $\{6, 8\} \in V(3)$; edge $(4, 5)$ is added since 4 and 5 are parents of 7 and 8 respectively where $\{7, 8\} \in V(3)$; edge $(7, 9)$ is added since 7 and 9 are parents of $11 \in V(4)$. The moralised graph can now be used to deduce the global Markov properties. In particular, it is clear that $3 \perp\!\!\!\perp 10 \mid 7$ since all paths from 3 to 10 pass through 7, whereas it is neither true that $1 \perp\!\!\!\perp 5 \mid (2, 4, 6)$ nor $9 \perp\!\!\!\perp 7 \mid (8, 11)$ since there are the 2 paths $(1, 5)$ and $(7, 9)$.

4.5 Granger Causality and Conditional Independence.

Consider the time series $\{V_k\}_{k \geq 1}$, $\{U_k\}_{k \geq 1}$ and $\{W_k\}_{k \geq 1}$. Causation between processes have been studied quite extensively in recent years. Granger's (1969) original definition of causality required that, for all time t , a best linear estimate of V_t based on $\{U_k\}_{k < t}$ and $\{V_k\}_{k < t}$, need only depend on $\{V_k\}_{k < t}$. A graphical representation of Granger causality is given by Lauritzen (1989). Here, however, the stronger definition of Florens and Mouchart (1985) is used — they say that

U is a *non-cause* of V , if for all time t :

$$V_t \perp \{U_k\}_{k < t} \mid \{V_k\}_{k < t}.$$

Thus the original idea of Granger is then replaced with a definition based on conditional independence itself and now the best estimate (rather than the best linear estimate) of V_t based on $\{U_k\}_{k < t}$ and $\{V_k\}_{k < t}$ need only depend on $\{V_k\}_{k < t}$. Although causality defined in terms of conditional independence does not quite capture what is usually meant by causality (Holland, 1986), non-causality of X on Y can often reasonably be considered as a consequence of there being a lack of causal relation from X to Y .

The class of models discussed in chapters 5 and 6 lead naturally to a definition of *conditional causality* which says that U is a *non-cause* of V *given* W if for all points in time t :

$$V_t \perp \{U_k\}_{k < t} \mid \{V_k\}_{k < t}, \{W_k\}_{k \leq t}.$$

Notice that a best estimate of V_t not only now depends on the past series $\{V_k\}_{k < t}$ and $\{W_k\}_{k < t}$ but also the current value W_t . Thus in any influence diagram of U_t , V_t and W_t satisfying the above definition of conditional causality, $W_t \in P(V_t)$ and $U_t \in Q(V_t)$. This concept of conditional causality is central to the theory of 2 new classes of Bayesian forecasting model introduced in the next two chapters.

4.6 Influence Diagram of the DLM.

The Bayesian forecasting system (see chapter 3) can be represented by an influence diagram. This section presents influence diagrams for a univariate time series $\{Y_k\}_{k > 0}$ following a DLM (see section 3.1) both before and after an observation is made at time t .

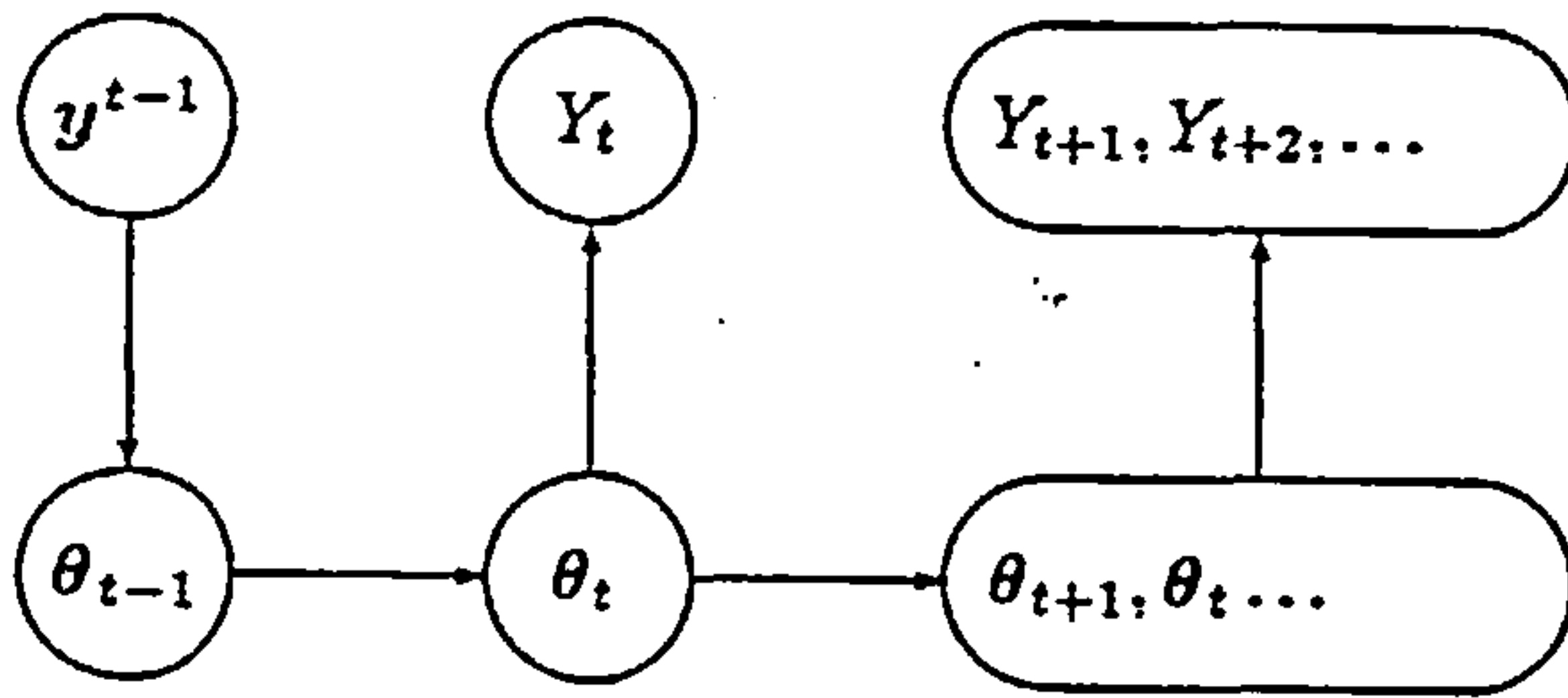


Figure 4.12: Influence diagram of DLM before observation y_t .

Figure 4.12 shows the influence diagram of the DLM before observation y_t is made. Notice that all information about previous observations y^{t-1} is contained in the posterior distribution of $\theta_{t-1} | y^{t-1}$. The prior distribution $\theta_t | y^{t-1}$ only depends on the posterior $\theta_{t-1} | y^{t-1}$ — this is why there are no edges between the node y^{t-1} and the nodes θ_t and θ_{t+1}, \dots in figure 4.12. Note also that the prior (forecast) distribution $Y_t | y^{t-1}$ only depends on the prior distribution $\theta_t | y^{t-1}$ and so there are no edges between the node y^{t-1} and the nodes Y_t and $Y_{t+1} \dots$ or the node θ_{t-1} and the nodes Y_t and $Y_{t+1} \dots$ in figure 4.12. By similar arguments, there are no edges between the nodes Y_t and $Y_{t+1} \dots$, Y_t and $\theta_{t+1} \dots$ or θ_t and $Y_{t+1} \dots$.

Figure 4.13 shows the influence diagram of the DLM after observation y_t has been made. Notice that edge (θ_t, Y_t) has now been reversed after the realisation of Y_t . This is because once the observation y_t has been made, the distribution of θ_t is updated. This arc reversal induces the edge (y^{t-1}, y_t) . The posterior distribution $\theta_t | y^t$ is then all that is required to find the prior distribution $\theta_{t+1} | y^t$. This suggests that the edge (θ_{t-1}, θ_t) should be reversed which in turn induces the edge (y^{t-1}, θ_t) . The nodes y^{t-1} and y_t can be combined into a single node y^t , then the node θ_{t-1} can be dropped from the influence diagram as it has no effect on the forecast of $Y_{t+1} \dots$. The resulting influence diagram is the same as in figure 4.12 one time point on.

PAGE

NUMBERING

AS ORIGINAL

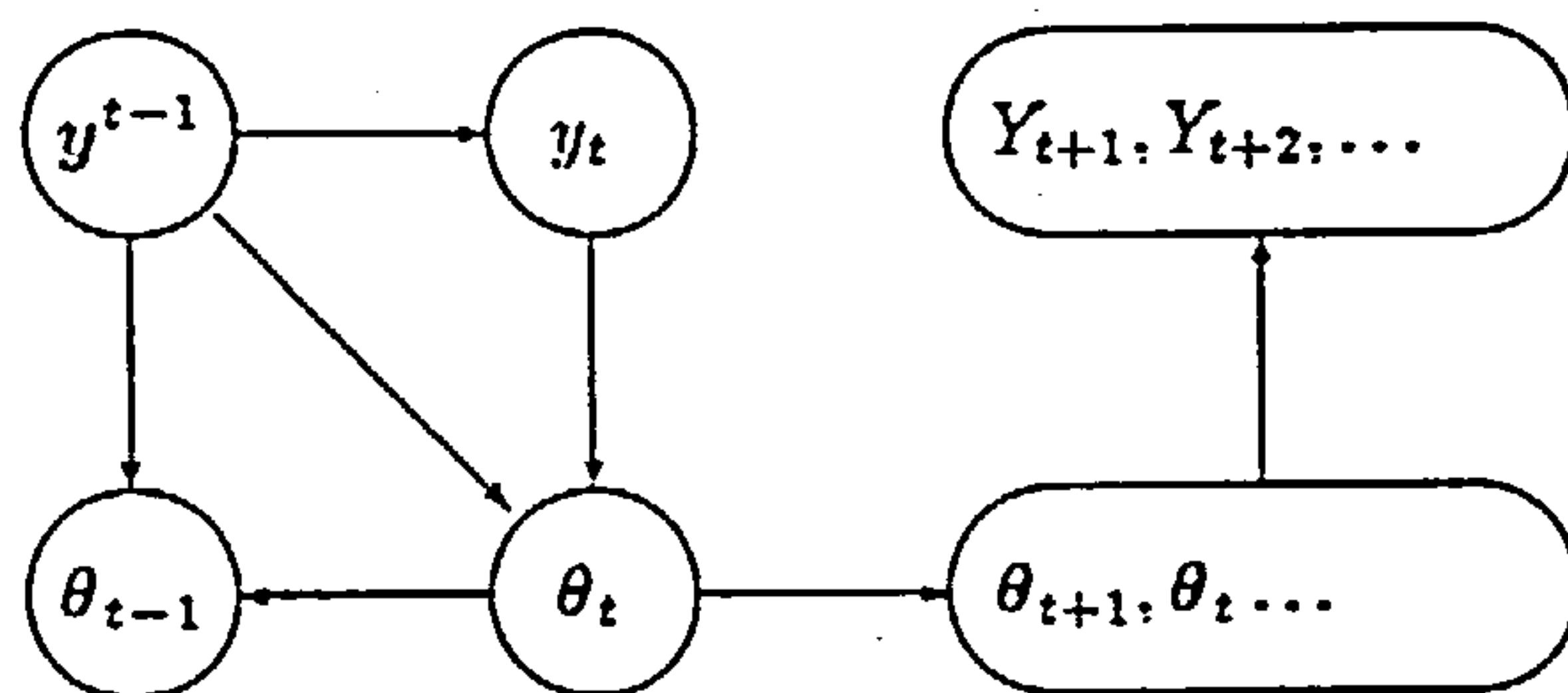


Figure 4.13: Influence diagram of DLM after observing y_t .

Chapter 5

Multiregression Dynamic Models.

5.1 Introduction.

This chapter introduces a new class of models called multiregression dynamic models (MDM's) in which graphical models and univariate Bayesian forecasting techniques are combined.

Let the general n -dimensional multivariate time series at time t be denoted by $Y_t^T = (Y_t(1), \dots, Y_t(n))$. Notice that this notation is slightly different from that introduced in chapter 3. Once again the observed value for $Y_t(r)$ is denoted by $y_t(r)$ and $y^t(r)^T = (y_1(r), \dots, y_t(r))$. For notational convenience set:

$$\begin{aligned} X_t(r)^T &= (Y_t(1), Y_t(2), \dots, Y_t(r-1)) \\ Z_t(r)^T &= (Y_t(r+1), \dots, Y_t(n)) \end{aligned} \tag{5.1}$$

In the symmetric model of Harvey (1986) and the DMR model (see section 3.6) each component of the series has the property that the forecast of each variable $Y_t(r)$, for $r = 1, \dots, n$, only depends on the past of that series so that:

$$Y_t(r) \Pi x^{t-1}(r), z^{t-1}(r) | y^{t-1}(r).$$

From the definition of non-causality introduced in section 4.5, this implies that the other components $\{X(r), Z(r)\}$ are non-causes of $Y(r)$. Thus, as Harvey (1989) points out, although the forecast covariances are easily estimated in these

systems, the price is the lack of any causal structure in the system. MDM's have been developed so that the causal structures which can be found in partially segmented markets are accommodated, whilst the model is also amenable to practical implementation where the forecast covariances are straightforward to calculate.

As MDM's have been developed with the accommodation of partially segmented markets in mind, the heuristic motivation provided by partially segmented markets for these models will be discussed in section 5.2. The MDM is formally defined in section 5.3 and proofs of the main results of this section are presented in section 5.4. Two special cases of the MDM which are especially simple to work with are studied in section 5.5 and full working examples of these two models are presented. Finally, section 5.6 provides a discussion of some of the properties of MDM's, their advantages and their limitations.

5.2 Heuristic Motivation for MDM's Provided by Partially Segmented Markets.

Once again consider the simple partially segmented market introduced in section 2.2 in which there are 3 brands X , Y and Z such that their segmentation can be represented heuristically by the undirected graph given in figure 2.1.

Now, suppose that there is a research hypothesis (Wermuth & Lauritzen, 1990) concerning certain causal relationships between the brand sales so that the sales of brand X can be considered as a causal factor in determining Y 's sales and the sales of brand Y can be considered as a causal factor in determining Z 's sales. For example, suppose that there are two types of consumer T_1 and T_2 so that $\{X, Y\} \in B_{T_1}$ and $\{Y, Z\} \in B_{T_2}$. Suppose that X has an aggressive promotion. Now this will be expected to have one of two possible effects:

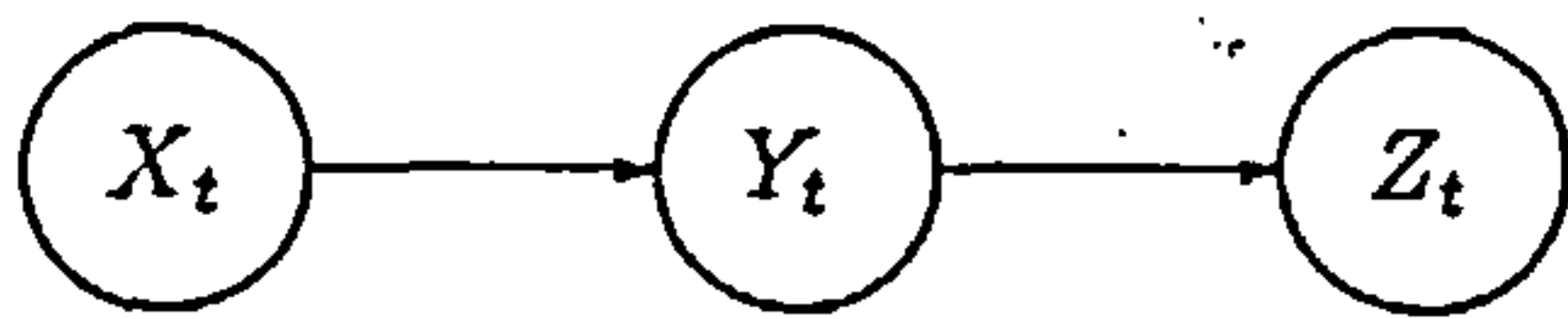


Figure 5.1: Heuristic influence diagram representing causal and conditional independence relationships at a fixed time t in the simplest possible partially segmented market.

1. X sells a lot of its brand which leads to a drop in Y 's sales and so there is an excess stock of brand Y . Thus X 's promotion will have an indirect effect on Z 's sales through the effect on Y 's sales.
2. Y quickly retaliates so that the sales of X and Y stabilise. Once again this will have an indirect effect on Z 's sales.

Wermuth & Lauritzen (1990) *heuristically* argued that when dealing with mixed graphical models, variables which are hypothesised to be causally linked should be connected by directed edges consistent with the direction of causality. The same argument is used here so that the undirected graph of figure 2.1 is used to incorporate the causality in the system so that a heuristic influence diagram is created representing the conditional independence *related to causality* between the brand sales at a fixed time period t . The heuristic graph of the influence diagram consistent with the causal relationships in the market is given in figure 5.1.

Suppose that the same influence diagram always represents the variables at all time points so that $P\{Y_t\} = X_t$ and $P\{Z_t\} = Y_t$ for $t \geq 1$. Suppose further that the processes over all time points up to and including time t can be represented by the graph of the influence diagram of figure 5.2 such that

$$P\{X_t\} \subseteq \{X^{t-1}\}$$

$$P\{Y_t\} \subseteq \{X^t, Y^{t-1}\}$$

$$P\{Z_t\} \subseteq \{Y^t, Z^{t-1}\}$$

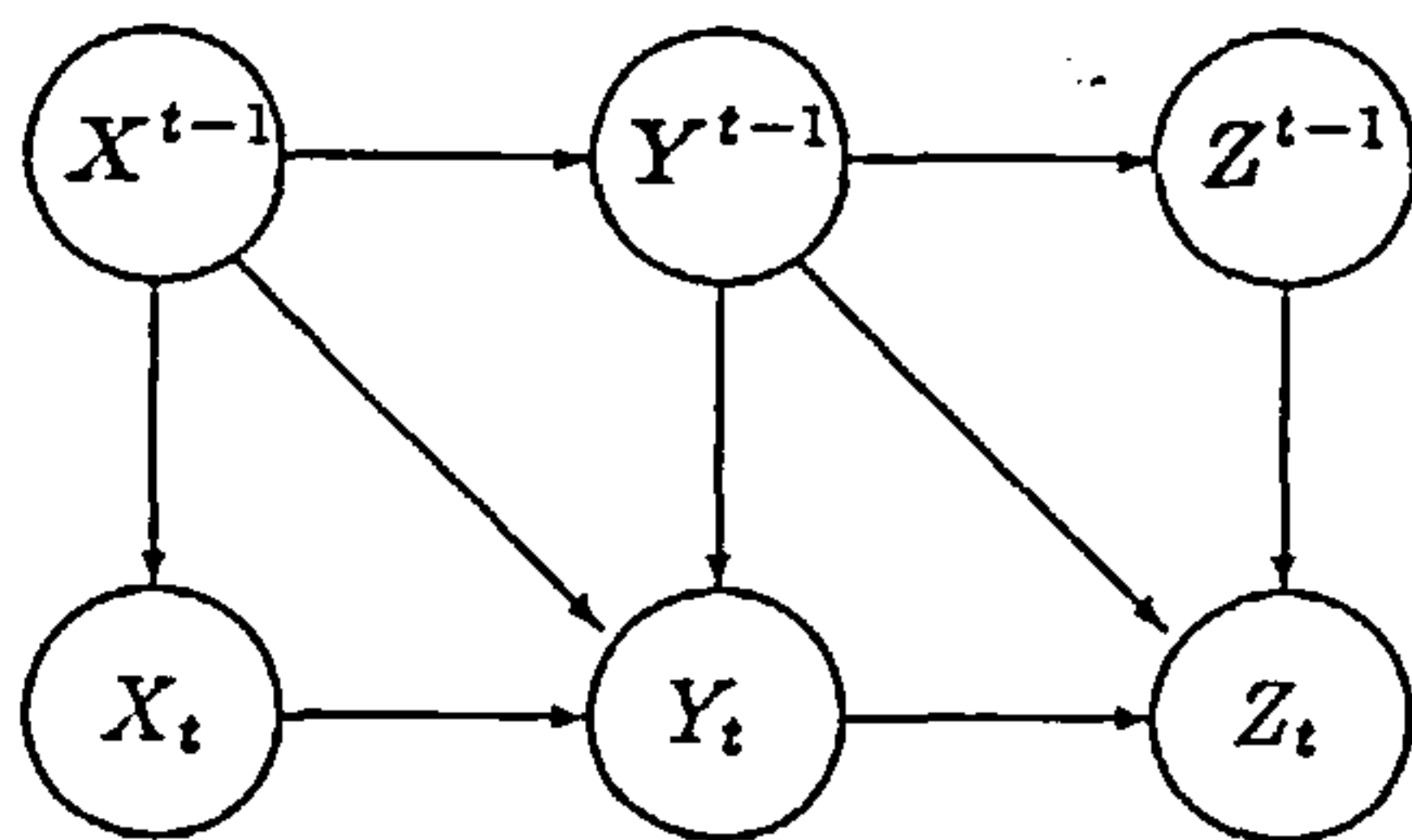


Figure 5.2: Graph of an ID representing the brand sales $\{X_k\}_{k \leq t}$, $\{Y_k\}_{k \leq t}$ and $\{Z_k\}_{k \leq t}$.

That is, each variable at time t has its past series as parents and also has the past series of its parents at time t , as parents. Notice that since there is never an edge (X_k, Z_k) , $1 \leq k \leq t-1$, there is no edge (X^{t-1}, Z^{t-1}) in figure 5.2.

These processes can be thought of as exhibiting conditional (Granger) causality, as introduced in section 4.5, such that Z is a non-cause of Y given X . Thus the best estimate of Y_t based on $\{X_k\}_{k \leq t}$, $\{Y_k\}_{k < t}$ and $\{Z_k\}_{k \leq t}$ need only depend on $\{Y_k\}_{k < t}$ and $\{X_k\}_{k \leq t}$. The conditional independences *related to causality* from figure 5.2 are such that:

$$Y_t \perp\!\!\!\perp \{Z_k\}_{k < t} \mid \{Y_k\}_{k < t}, \{X_k\}_{k \leq t}.$$

Thus, given that the variables exhibit these conditional causal relationships, an appropriate forecasting model for each variable at time t , is simply a function of its own past series, the past series of its parents and the value of its parents *at time t*. For example, the general form of an appropriate forecasting model for the series $\{Y_k\}_{k \geq 1}$ is given by:

$$Y_t = f(x^t, y^{t-1}, \theta_t) + v_t,$$

where $f(\cdot)$ is some function and v_t has some probability distribution. This reasoning can be generalised directly to the case where there are n series.

Suppose that for fixed time t , the structure of the n -dimensional multivariate series Y_t can be heuristically represented by an influence diagram I such that:

$$P\{Y_t(r)\} \subseteq X_t(r).$$

Suppose further that the process $\{Y_k\}_{k \leq t}$ can be heuristically represented by an influence diagram I^* such that:

$$P\{Y_t(r)\} \subseteq \{X^t(r), Y^{t-1}(r)\}.$$

Once again it can be seen that for the vector series $\{Y_k\}_{k \geq 1}$, $Z(r)$ can be considered as a non-cause of $Y(r)$ given $X(r)$ so that:

$$Y_t(r) \amalg Z^{t-1}(r) \mid \{Y^{t-1}(r), X^t(r)\}. \quad (5.2)$$

Let $\theta_t^T = (\theta_t(1)^T, \dots, \theta_t(n)^T)$ be the state vectors determining the distributions of $Y_t(1), Y_t(2), \dots, Y_t(n)$ respectively, and let s_r be the dimension of the vector $\theta_t(r)$, $t \geq 1$. An appropriate forecasting model for each $Y_t(r)$ consistent with this non-causality is then of the form

$$Y_t(r) = f(x^t(r), y^{t-1}(r), \theta_t(r)) + v_t(r), \quad r = 1, \dots, n \quad (5.3)$$

where $f(\cdot)$ is some function and $v_t(r)$ has some probability distribution.

The MDM uses this methodology for deriving its observation equations and through the special form of the system equation breaks a complex multivariate problem into n univariate ones. Notice how the causal relationships between the variables are accommodated and that the stringent symmetry conditions imposed on the variables in the models of Harvey (1986) and the DMR model (see section 3.6) are not required. The MDM will now be formally defined.

5.3 The Multiregression Dynamic Model.

Using the notation introduced for the n -dimensional vector time series at the beginning of this chapter, call $\{Y_t\}_{t \geq 1}$ a *Multiregression Dynamic Model* (MDM) if it is governed by the following n observation equations, system equation and initial information where the restrictions on these equations given below hold for all points in time:

Observation equations

$$Y_t(r) = F_t(r)^T \theta_t(r) + v_t(r) \quad v_t(r) \sim (0, V_t(r)), \quad 1 \leq r \leq n \quad (5.4)$$

System equation

$$\theta_t = G_t \theta_{t-1} + w_t \quad w_t \sim (0, W_t) \quad (5.5)$$

Initial information

$$(\theta_0 | D_0) \sim (m_0, C_0) \quad (5.6)$$

The s_r dimensional column vector $F_t(r)$ is allowed to be an arbitrary but known function of $x^t(r)$ and $y^{t-1}(r)$, but not $z^t(r)$ and $y_t(r)$; $V_t(1), \dots, V_t(n)$ are the known scalar observation variances; the $(s \times s)$ matrices

$$G_t = \text{blockdiag}(G_t(1), \dots, G_t(n)),$$

$$W_t = \text{blockdiag}(W_t(1), \dots, W_t(n))$$

and

$$C_0 = \text{blockdiag}(C_0(1), \dots, C_0(n))$$

are assumed known and are such that $G_t(r)$, $W_t(r)$ and $C_0(r)$ are $(s_r \times s_r)$ square matrices which may be functions of past vectors $x^{t-1}(r)$ and $y^{t-1}(r)$ but nothing else.

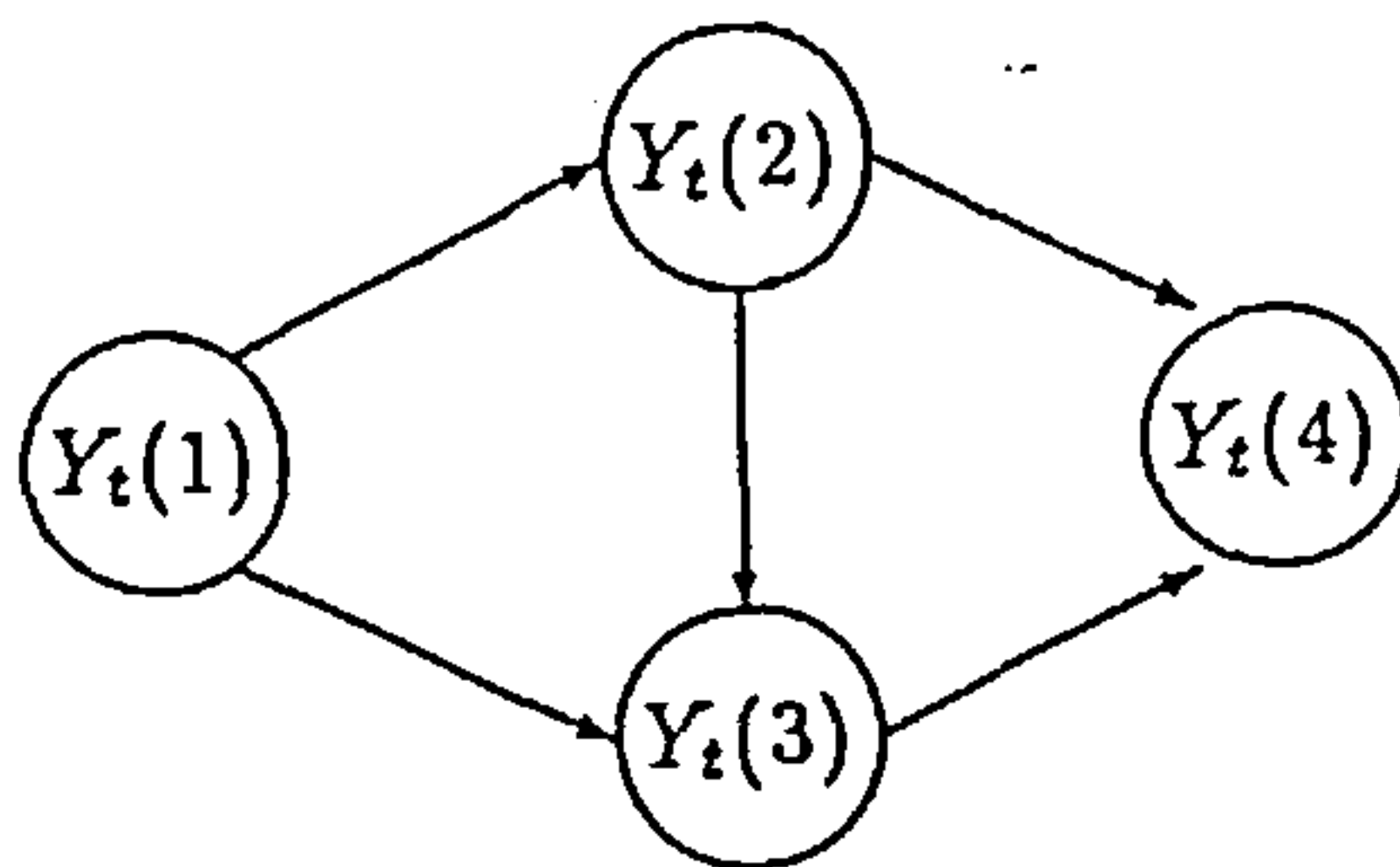


Figure 5.3: Graph of the ID describing $Y_t(1), \dots, Y_t(4)$ to illustrate how an MDM is derived.

The error vectors

$$\mathbf{v}_t^T = (v_t(1), \dots, v_t(n))$$

and

$$\mathbf{w}_t^T = (\mathbf{w}_t(1)^T, \dots, \mathbf{w}_t(n)^T),$$

where $v_t(r)$ is the observation error and $\mathbf{w}_t(r)$ is the s_r dimensional error vector for $Y_t(r)$, $1 \leq r \leq n$, are such that variables $v_t(1), \dots, v_t(n)$ and $\mathbf{w}_t(1), \dots, \mathbf{w}_t(n)$ are all mutually independent and the vectors $\{\mathbf{v}_t, \mathbf{w}_t\}_{t \geq 1}$ are mutually independent with time.

To illustrate how an MDM would be defined for a vector time series, consider the following example. Suppose that $\mathbf{Y}_t^T = (Y_t(1), \dots, Y_t(4))$ where at any fixed time t \mathbf{Y}_t can be represented by the graph of the influence diagram given in figure 5.3. The MDM for \mathbf{Y}_t would have observation equations in which $F_t(1)$ is a function of $\mathbf{y}^{t-1}(1)$; $F_t(2)$ is a function of $\{\mathbf{y}^t(1), \mathbf{y}^{t-1}(2)\}$; $F_t(3)$ is a function of $\{\mathbf{y}^t(1), \mathbf{y}^t(2), \mathbf{y}^{t-1}(3)\}$; and $F_t(4)$ is a function of $\{\mathbf{y}^t(2), \mathbf{y}^t(3), \mathbf{y}^{t-1}(4)\}$.

There are two important results which are central to the theory of MDM's, the proofs of which will be given in the next section. It is shown that if

$$\prod_{r=1}^n \theta_{t-1}(r) | \mathbf{y}^{t-1}$$

then the following conditional independence statements must hold:

Result 1.

$$\prod_{r=1}^n \theta_t(r) | y^t \quad (5.7)$$

In words this means that if $\{\theta_{t-1}(r)\}$ are mutually independent given the data y^{t-1} , then under the DLM $\{\theta_t(r)\}$ are also independent given y_t . It will follow by induction that, provided $\{\theta_0(r)\}$ are initially independent, the parameters remain independent for all time given the current available information.

Result 2.

$$\theta_t(r) \prod z^t(r) | x^t(r), y^t(r) \quad (5.8)$$

Written in terms of the components of y^t this reads:

$$\theta_t(r) \prod y^t(r+1), y^t(r+2), \dots, y^t(n) | y^t(1), \dots, y^t(r)$$

In words this means that, given the past observation vectors of the first r indexed series, $\theta_t(r)$ is independent of the rest of the past data.

When defining the MDM, C_0 is set to be block diagonal and so the parameters for each variable are initially mutually independent. Therefore, by Result 1, the parameters associated with each variable are updated independently after each observation and remain independent at each time point. Thus, as each variable follows a conditional univariate Bayesian dynamic model, the conditional distribution for each variable can be updated independently and conditional forecasts can be found separately. A complex multivariate problem has therefore been decomposed into n univariate ones. Notice that no assumption of normality has been made for both Results 1 and 2 and when defining the MDM and so the class of MDM is extremely large. However, even when $F_t(r)$ is a non-linear function of $\{x^t(r), y^{t-1}(r)\}$, because it is assumed known and therefore fixed at time t , a *normal MDM* can be defined such that v_t and w_t are Gaussian so that, *conditional on* $x_t(r)$, $Y_t(r)$ is Gaussian. The MDM is particularly simple to work

with in this case. Each variable follows a normal DLM and as such, updating and forecasts of conditional univariate distributions are identical to those presented in section 3.1. Section 5.5 discusses two special cases of normal MDM's called the *linear multiregression dynamic model* (LMDM) and the *corrected linear multiregression dynamic model* (CLMDM) in which $F_t(r)$ is a linear function of $\{x^t(r), y^{t-1}(r)\}$.

Of course, in reality, the series are observed simultaneously so that $x_t(r)$ will not be observed before a forecast for $Y_t(r)$ is made. The marginal forecast distributions for each variable are therefore required. In general these marginal distributions will not be Gaussian. However, the moments of Y_t for many MDM's can be derived fairly easily. The derivation of the first two moments of Y_t for the CLMDM and the first moment of the LMDM are illustrated in section 5.5.

Results 1 and 2 combine to enable the joint one-step ahead forecast distribution of Y_t to be simplified. Consider the forecast distribution for n brands:

$$p(y_t | y^{t-1}) = \int_{\theta_t} p\{y_t | \theta_t, y^{t-1}\} p\{\theta_t | y^{t-1}\} d\theta_t$$

The conditional independence statements of Result 1, together with the structure of the vector $F_t(r)$ specified by the MDM, ensure that the joint distribution can be expressed as the product of the individual forecast distributions of $y_t(r)$ with regressors $x_t(r)$. So:

$$\begin{aligned} p(y_t | y^{t-1}) &= \prod_r p\{y_t(r) | x^t(r), y^{t-1}(r)\} \\ &= \prod_r \int_{\theta_{t(r)}} p\{y_t(r) | x^t(r), y^{t-1}(r), \theta_{t(r)}\} p\{\theta_{t(r)} | y^{t-1}\} d\theta_{t(r)} \end{aligned}$$

However, by the conditional independence statements of Result 2 and the specified structure of $G_t(r)$, $p(\theta_{t(r)} | y^{t-1})$ only depends on $x^{t-1}(r)$ and $y^{t-1}(r)$ and

thus can be rewritten as:

$$p \{ \theta_t(r) | y^{t-1} \} = p \{ \theta_t(r) | x^{t-1}(r), y^{t-1}(r) \}$$

Before a discussion of MDM's is presented, the next section formally proves Results 1 and 2.

5.4 Formal presentation of the results.

Theorem 4.2.1 from section 4.2 can be directly used to prove the following theorem. For convenience the following notation will be used:

$$\phi_t(r)^T = (\theta_t(1)^T, \dots, \theta_t(r-1)^T)$$

$$\psi_t(r)^T = (\theta_t(r+1)^T, \dots, \theta_t(n)^T)$$

Theorem 5.4.1 *Let $\{Y_t\}_{t \geq 1}$ be governed by an MDM and, using the notation of the previous section, assume:*

$$\phi_{t-1}(r) \perp\!\!\!\perp y^{t-1}(r), z^{t-1}(r) | x^{t-1}(r) \quad r = 2, \dots, n \quad (5.9)$$

$$\theta_{t-1}(r) \perp\!\!\!\perp z^{t-1}(r), \phi_{t-1}(r) | x^{t-1}(r), y^{t-1}(r) \quad r = 1, \dots, n \quad (5.10)$$

$$\psi_{t-1}(r) \perp\!\!\!\perp \theta_{t-1}(r), \phi_{t-1}(r) | y^{t-1} \quad r = 2, \dots, n-1 \quad (5.11)$$

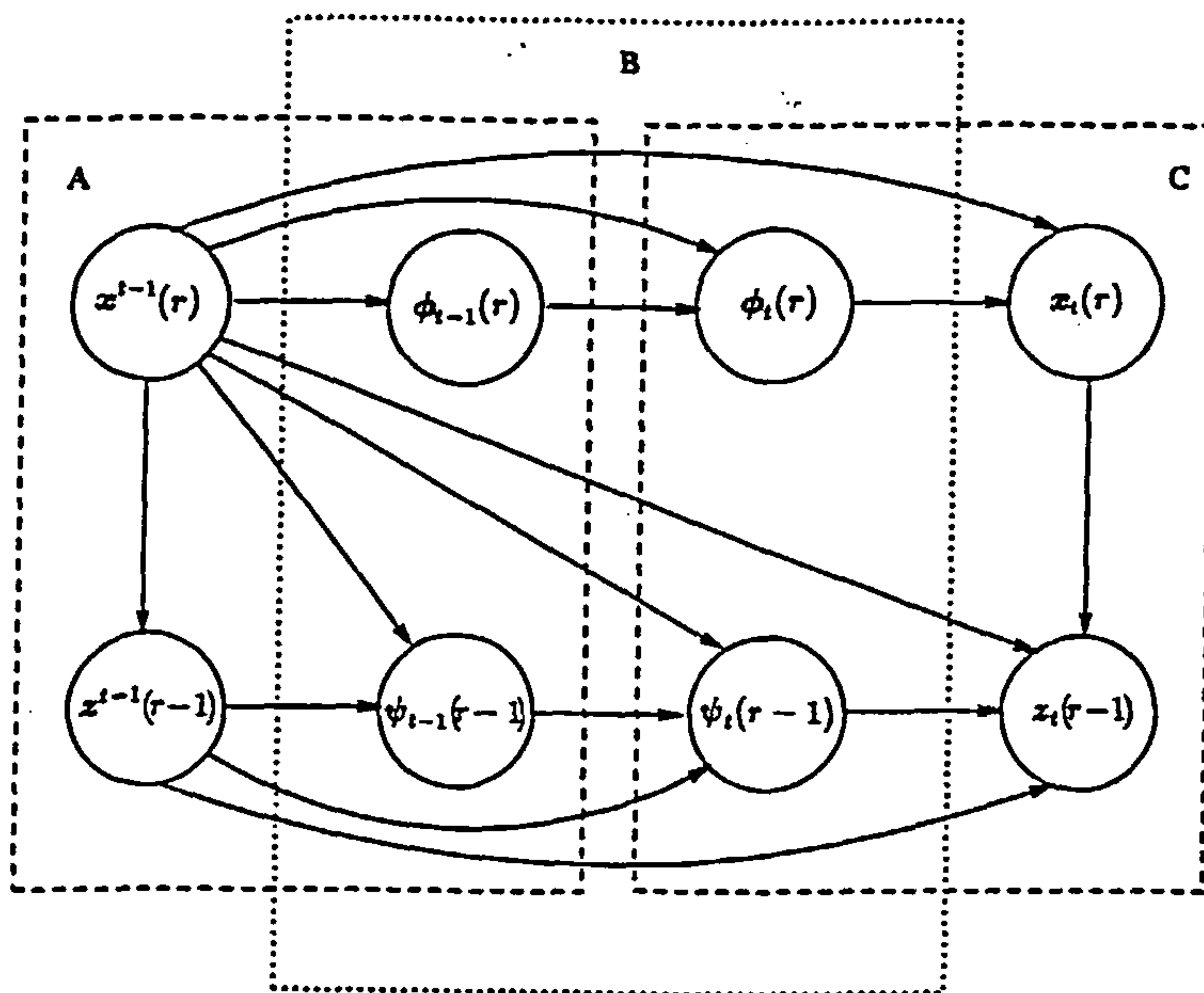
Then the following conditional independence statements must also be true:

$$\phi_t(r) \perp\!\!\!\perp y^t(r), z^t(r) | x^t(r) \quad r = 2, \dots, n \quad (5.12)$$

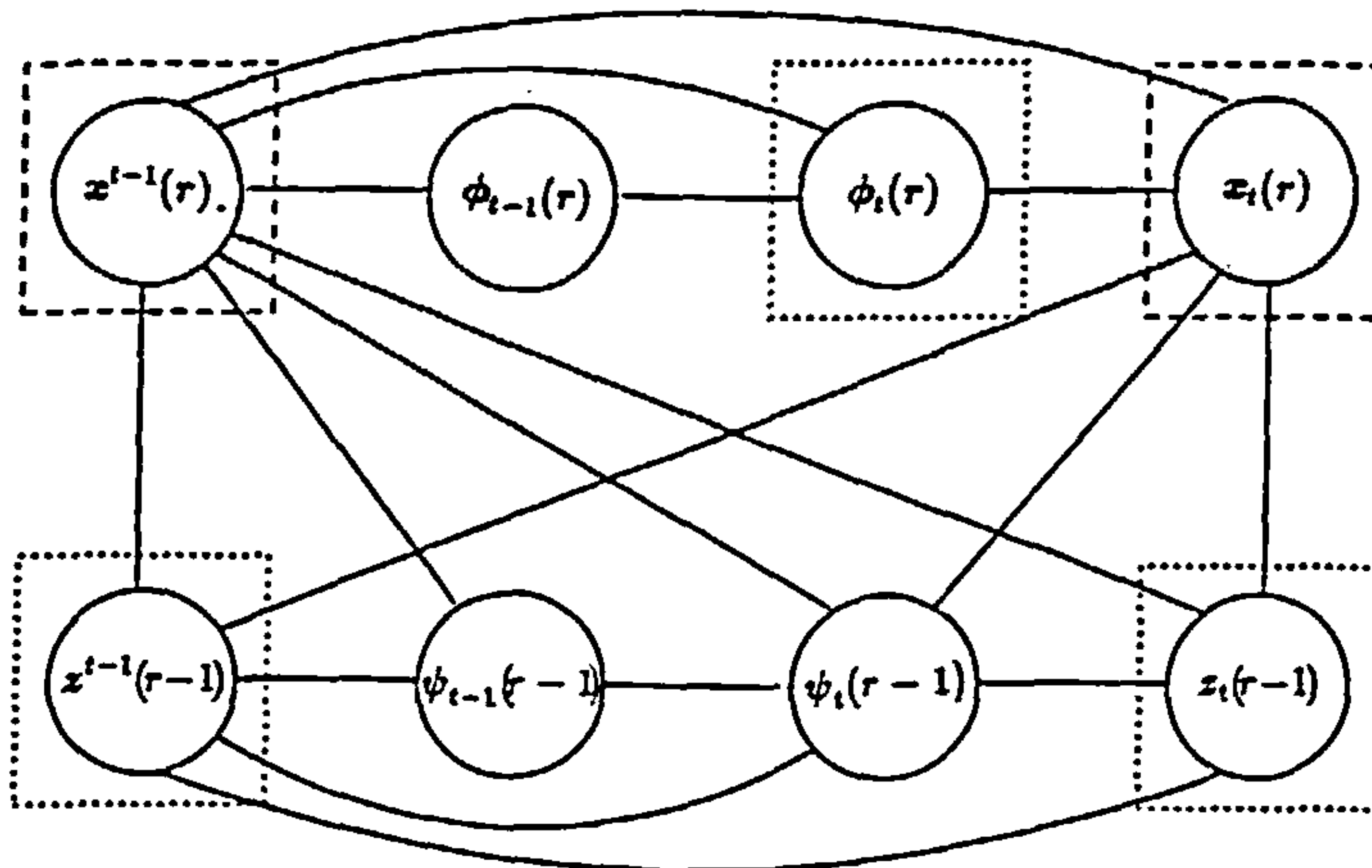
$$\theta_t(r) \perp\!\!\!\perp z^t(r), \phi_t(r) | x^t(r), y^t(r) \quad r = 1, \dots, n \quad (5.13)$$

$$\psi_t(r) \perp\!\!\!\perp \theta_t(r), \phi_t(r) | y^t \quad r = 2, \dots, n-1 \quad (5.14)$$

Proof: First the conditional independence statements contained in the inductive hypotheses 5.9, 5.10 and 5.11 respectively and the MDM itself, will be represented in the three graphs of influence diagrams I1, I2 and I3 of figures 5.4, 5.5 and 5.6.



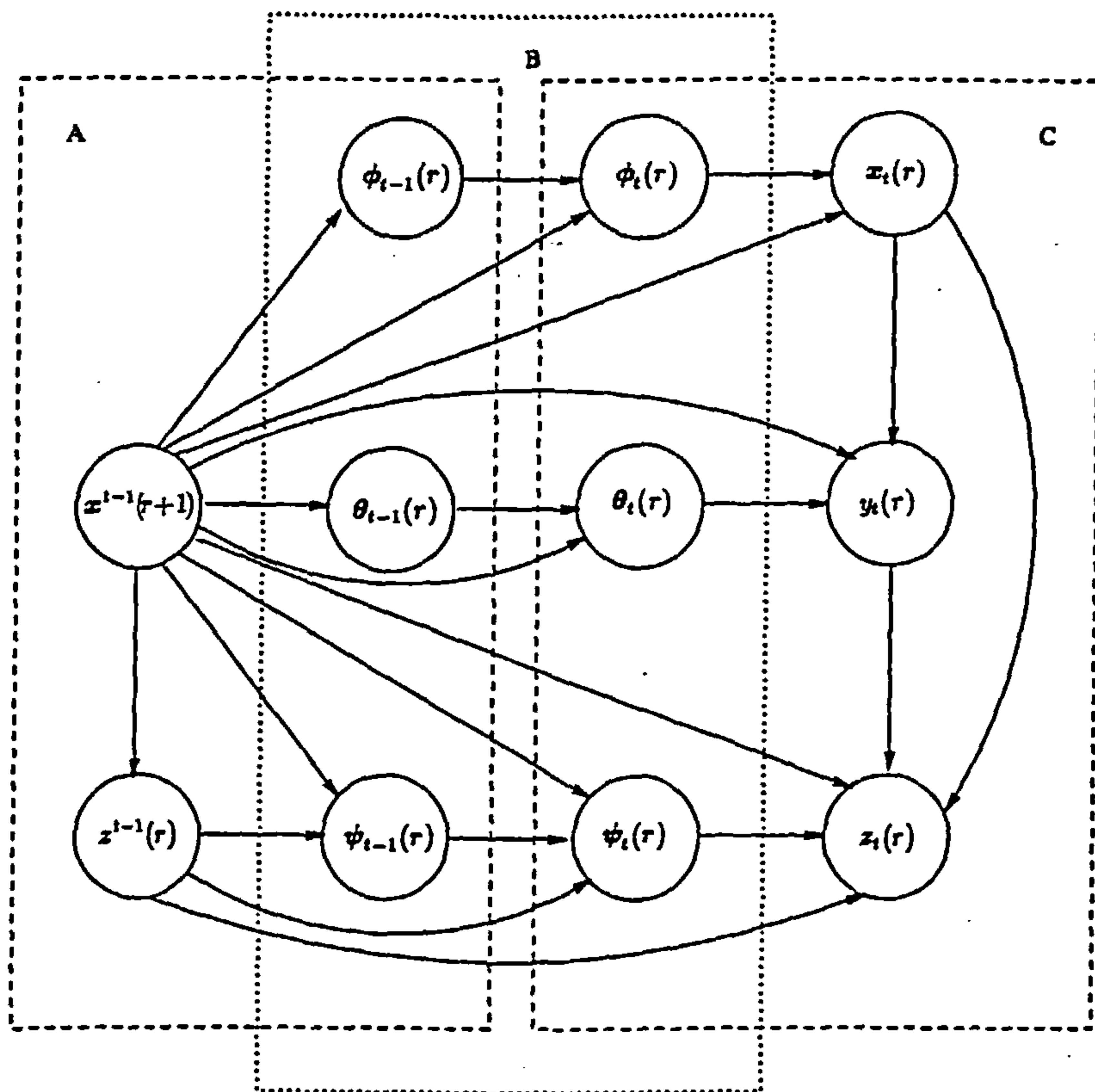
Graph I1



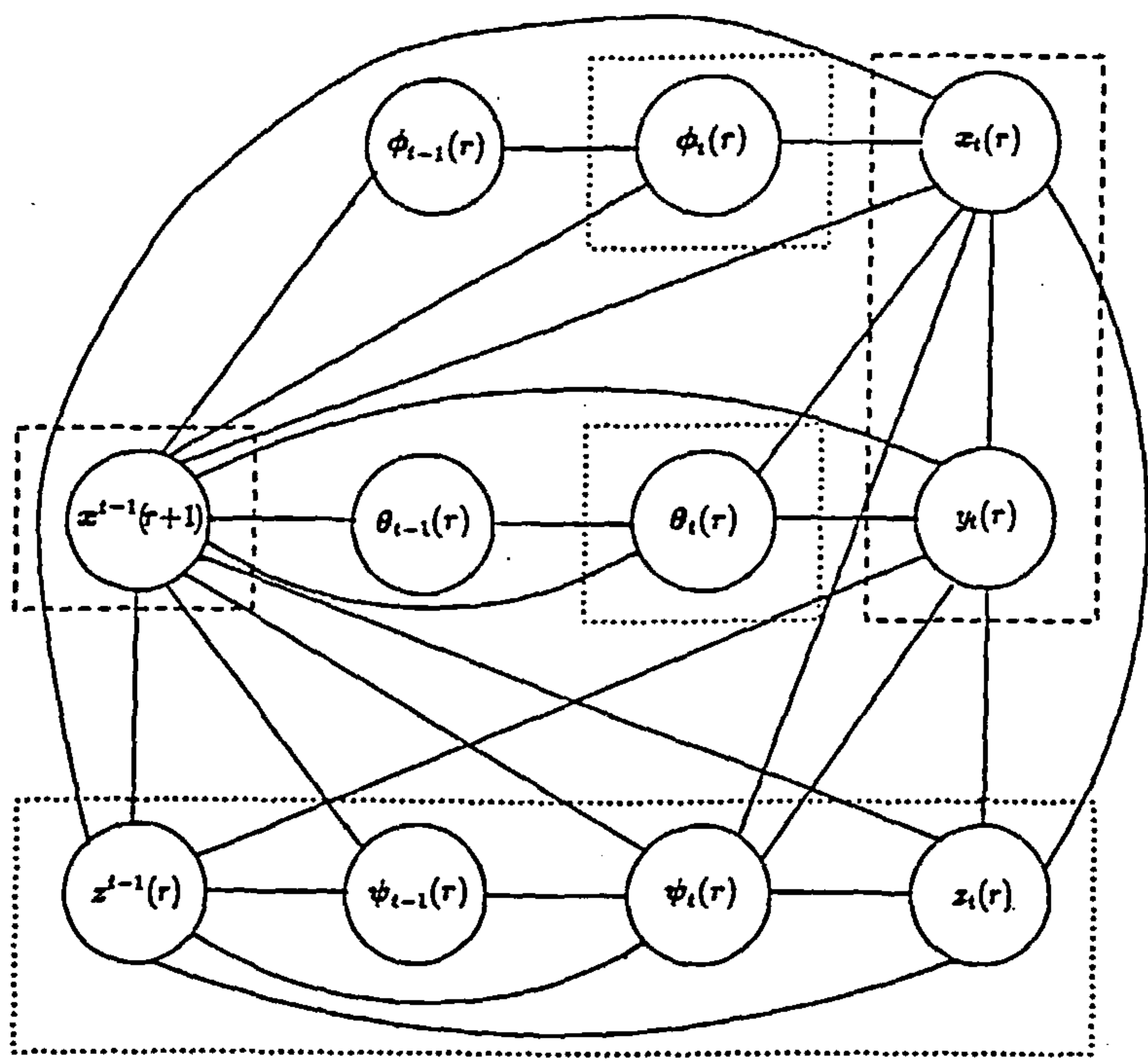
Graph J1

 = conditioning variables

Figure 5.4: The graph of the influence diagram I1 and its moralised graph J1.



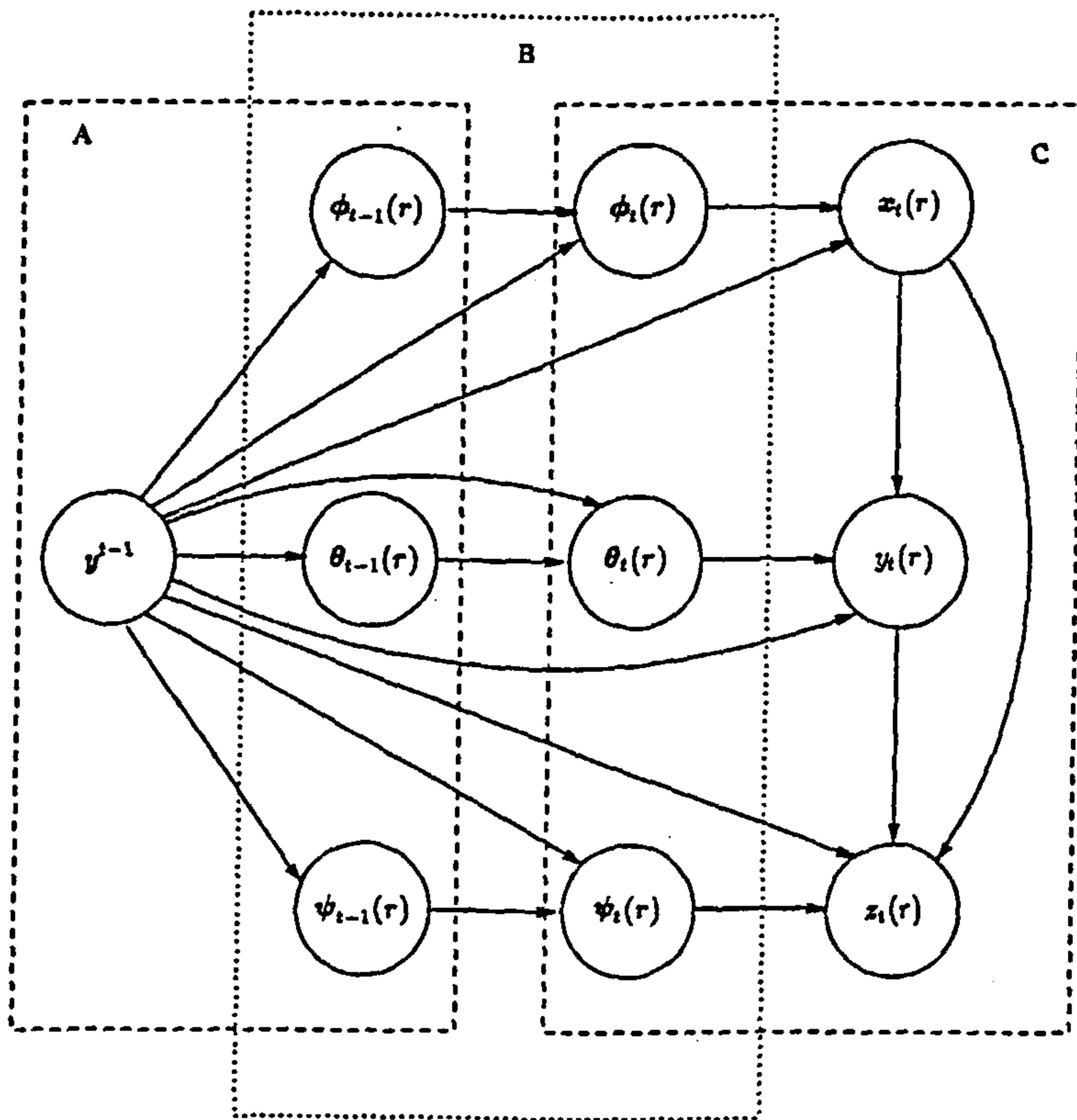
Graph I2



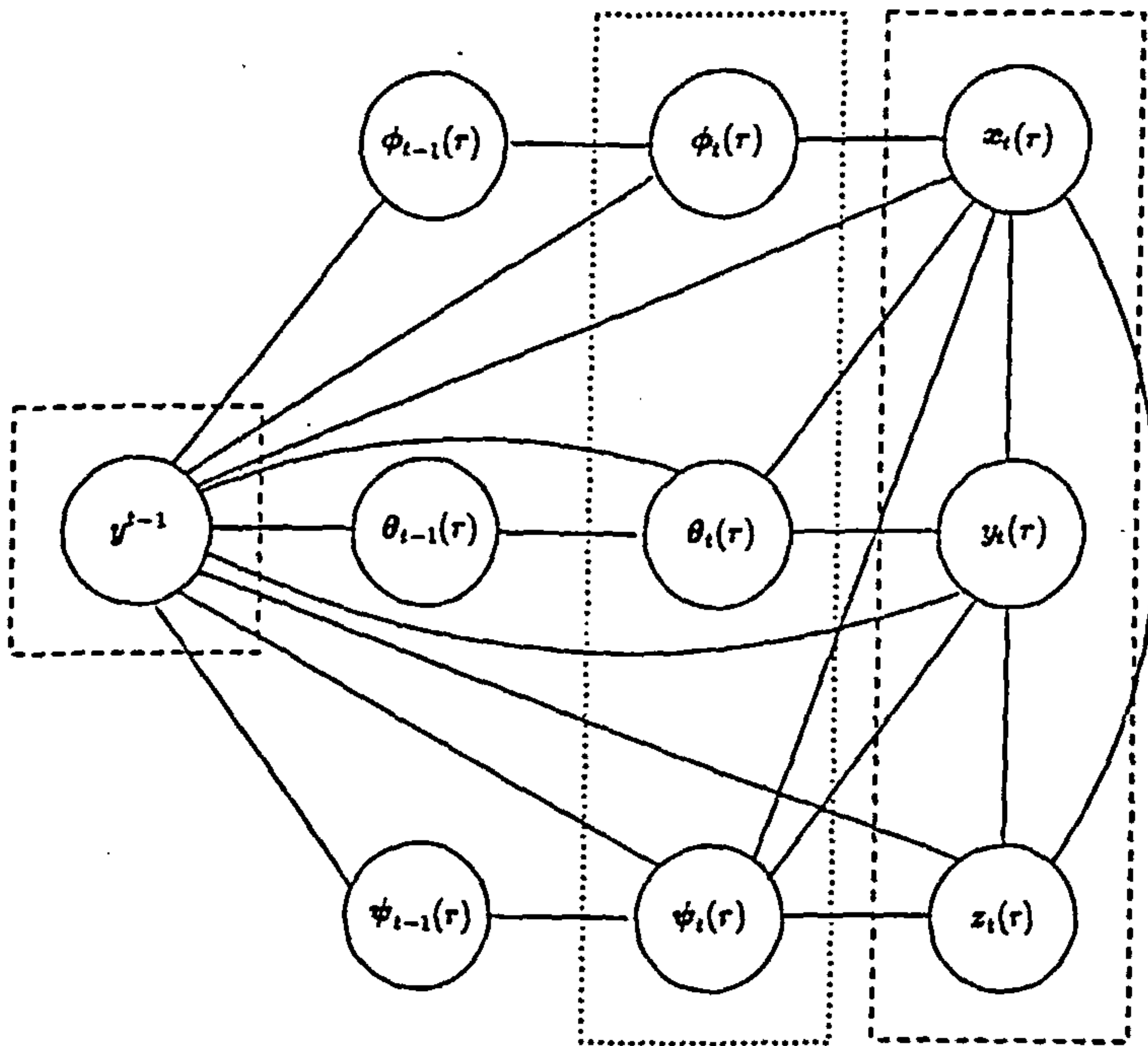
Graph J2

 = conditioning variables

Figure 5.5: The graph of the influence diagram I2 and its moralised graph J2.



Graph I3



Graph J3

 = conditioning variables

Figure 5.6: The graph of the influence diagram I3 and its moralised graph J3.

Using property *P3* of the introduction, assumption 5.10 can be equivalently stated by the pair of assertions

$$\theta_{t-1}(r) \perp\!\!\!\perp \phi_{t-1}(r) \mid y^{t-1} \quad (5.15)$$

$$\theta_{t-1}(r) \perp\!\!\!\perp z^{t-1}(r) \mid x^{t-1}(r), y^{t-1}(r) \quad (5.16)$$

Now, by assumption 5.9 and assertion 5.16, arcs between nodes of subvectors of y^{t-1} whose components all have an index in y_t greater than r to those nodes of subvectors of θ_t whose components have indices less than or equal to r , are allowed to be omitted.

On the other hand, assertion 5.15 and assumption 5.11 allow arcs between parameter vectors which do not share the same components to be omitted.

This justifies the implied conditional independence statements in the boxes labelled A in figures 5.4, 5.5 and 5.6.

The block diagonal form of G_t and W_t in the MDM ensure that:

$$\theta_t(r) \perp\!\!\!\perp \theta_{t-1} \setminus \{\theta_{t-1}(r)\} \mid y^{t-1}, \theta_{t-1}(r) \quad 1 \leq r \leq n$$

Thus on the graph of the influence diagram any arcs between sets of components of θ_{t-1} and sets of components of θ_t with no common index can be omitted. This argument justifies the omission of arcs in box B of the influence diagram boxes.

Finally the conditional independence statements implicit in the observation equations of the M.D.M mean that:

$$y_t(r) \perp\!\!\!\perp \theta_t \setminus \{\theta_t(r)\} \mid y^{t-1}(r), x_t(r), \theta_t(r)$$

These statements justify the omission of arcs in box C of the influence diagram graphs.

The moralised undirected graphs J1, J2 and J3 of the influence diagrams I1, I2 and I3 are shown in figures 5.4, 5.5 and 5.6 respectively. It is now simple to

check that the set of nodes representing the conditioning variables block all paths between the nodes associated with elements in different sets U and V claimed to be independent. So by the theorem 4.2.1, the result is proved.

The results introduced in section 5.3 can now be formally stated in the following corollary of theorem 5.4.1:

Corollary 5.4.2 *If in an MDM the initial states are independent, that is if $\Pi_{r=1}^n \theta_0(r)$, then for all time t :*

$$\Pi_{r=1}^n \theta_i(r) | y^t$$

and

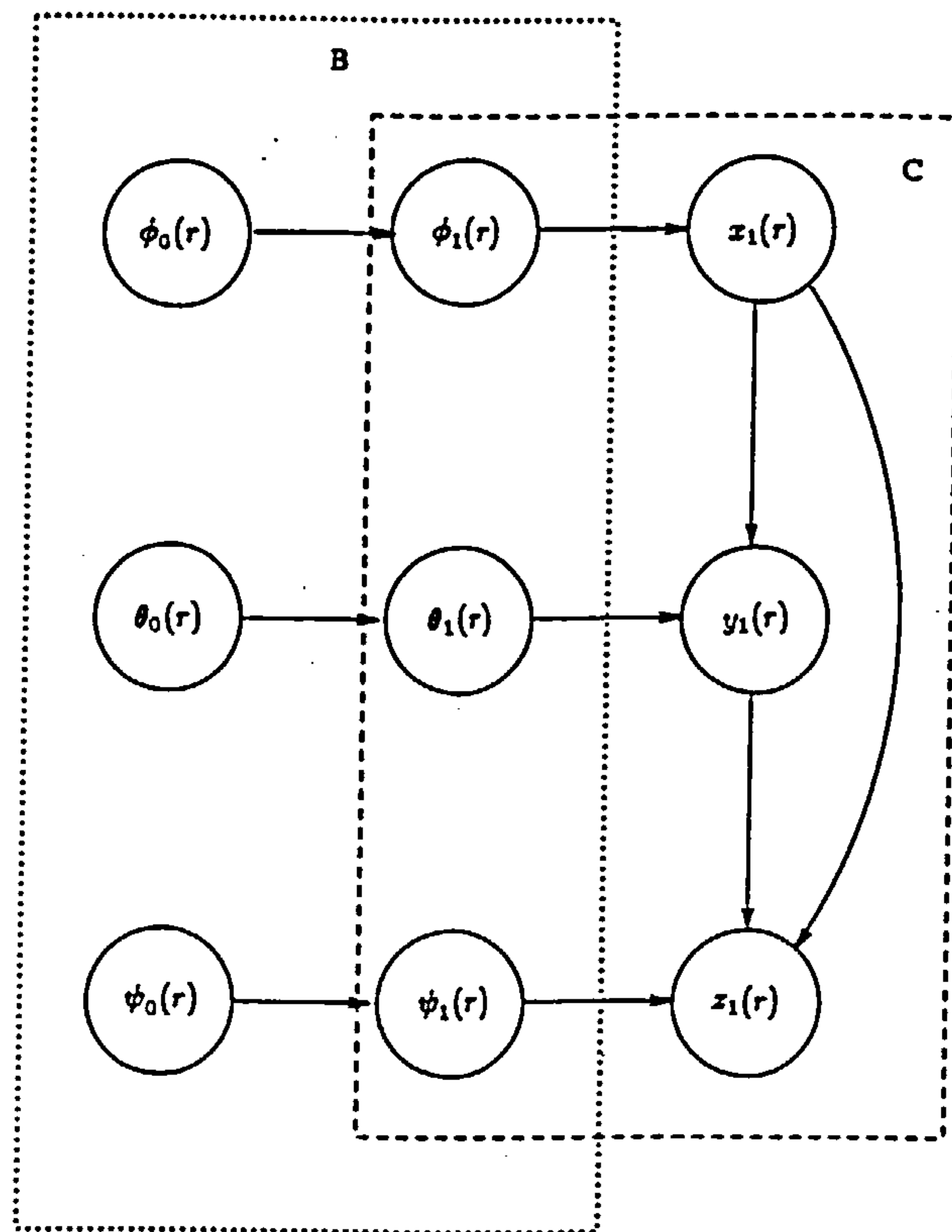
$$\theta_i(r) \Pi y^t(r+1), \dots, y^t(n) | y^t(1), \dots, y^t(r)$$

Proof: To prove the result for $t = 1$ proceed as in the theorem above. From the hypotheses (set A), the system equation (set B) and the observation equations (set C) the graph of the influence diagram I4 is obtained in figure 5.7.

By using exactly the same argument as in the theorem and drawing the corresponding moralised undirected graph J , the conditional independence statements 5.12, 5.13 and 5.14 of theorem 5.4.1 can now be deduced for time $t = 1$. Therefore, since $\Pi_{r=1}^n \theta_0(t)$ under the corollary, theorem 5.4.1 must be true for $t = 1$, and so by induction the assertions of theorem 5.4.1 must be true for all time t .

Since theorem 5.4.1 holds for all time t , the conditional independence statements from the theorem can be combined:

$$\theta_i(r) \Pi \{ \phi_i(r), \psi_i(r) \}, z^t(r) | x^t(r), y^t(r) \quad (5.17)$$



Graph I4

Figure 5.7: The graph of the influence diagram I4.

Now, by property $P3$, statement 5.17 implies that:

$$\theta_t(r) \quad \Pi \quad \{\phi_t(r), \psi_t(r)\} \mid z^t(r), x^t(r), y^t(r)$$

therefore,

$$\Pi_{r=1}^n \theta_t(r) \mid y^t$$

Statement 5.17 also implies that:

$$\theta_t(r) \quad \Pi \quad z^t(r) \mid x^t(r), y^t(r)$$

and so

$$\theta_t(r) \quad \Pi \quad y^t(r+1), \dots, y^t(n) \mid y^t(1), \dots, y^t(r)$$

therefore the corollary and hence the results of section 5.3 are proved.

These results apply to a very wide class of models. However it is helpful initially to consider the implications for the LMDM and CLMDM because the joint distribution of the variables in these models can be calculated explicitly and the one-step ahead forecasts have a particularly simple form.

5.5 Linear Multiregression Dynamic Models.

Consider the MDM defined by equations 5.4, 5.5 and 5.6. This section introduces two special cases of the MDM — the *linear MDM* (LMDM) and the *corrected linear MDM* (CLMDM) — which are especially simple to work with.

Suppose that $\{v_t(r), w_t(r)\}_{t \geq 1}$ are jointly Gaussian, are independent of y^{t-1} and $F_t(1)$ does not depend on Y_t . Now, let

$$F_t(r)^T = (x_t(r)^T, \tilde{x}_t(r)^T), \quad 2 \leq r \leq n,$$

where $x_t(r)^T = (y_t(1), \dots, y_t(r-1))$ and $\tilde{x}_t(r)$ is a set of known exogenous

variables not dependent upon $\mathbf{x}_t(r)$. This process will be known as a *Linear Multiregression Dynamic Model* (LMDM). Alternatively, suppose that

$$\mathbf{F}_t(r)^T = \left[\left\{ \mathbf{x}_t(r) - \hat{\mathbf{f}}_t(r) \right\}^T, \tilde{\mathbf{x}}_t(r)^T \right],$$

where

$$\hat{\mathbf{f}}_t(r)^T = \left[E \{ Y_t(1) | \mathbf{y}^{t-1} \}, \dots, E \{ Y_t(r-1) | \mathbf{y}^{t-1} \} \right] \quad (5.18)$$

and, from Result 2 of section 5.3, $\hat{\mathbf{f}}_t(r)$ only depends on \mathbf{y}^{t-1} through $\mathbf{x}^{t-1}(r)$ and $\mathbf{y}^{t-1}(r)$. Once again $\tilde{\mathbf{x}}_t(r)$ is a set of known exogenous variables not dependent on $\mathbf{x}_t(r)$. The process will then be known as a *Corrected Linear Multiregression Dynamic Model* (CLMDM). Thus given the components $\mathbf{x}_t(r)$ which precede it, each component $Y_t(r)$ is described as a *univariate* DLM with regressors contained in the vectors $\mathbf{F}_t(r)$ given above.

Note that the LMDM is a stochastic version of a recursive simultaneous equations model where $Y_t(r)$ is regressed against a subset of contemporary variables listed before $Y_t(r)$. Harvey (1989), considers a degenerate special case of this. Unlike the economic literature, however, the emphasis here is not on parameter estimation but on the types of predictive joint distribution that \mathbf{Y}_t can exhibit given the inevitable continued uncertainty about the regression state in the process.

The causal relationships between brands modelled by an LMDM have a different interpretation to those modelled by a CLMDM. Suppose that there are just two variables, $\{Y_k(1)\}_{k \geq 1}$ and $\{Y_k(2)\}_{k \geq 1}$, such that $Y(1)$ is thought to be a causal factor of $Y(2)$. If $\{Y_k\}_{k \geq 1}$ follows an LMDM, then a long term level change in $Y(1)$'s level would cause a sustained level change in $Y(2)$. However, if $\{Y_k\}_{k \geq 1}$ follow a CLMDM, then the same long term level change in $Y(1)$ would have a different causal effect on $Y(2)$. In this case the level change in $Y(2)$ would

be short term, followed by a drift back to the original level. This is due to the fact that the CLMDM uses residuals as regressors unlike the LMDM which uses the actual observation. Thus the CLMDM essentially relates causality through forecast residuals so that misforecasting of $Y_t(1)$ helps in adjusting the forecast distribution of $Y_t(2)$ while the reverse is not true. Therefore, a sudden level change in $Y(1)$ would lead to a large residual value in the model for $Y(2)$, thus leading to a level change in $Y(2)$. As the model for $Y(1)$ adapts to the level change, so the residual in the regression term in $Y(2)$'s model will get smaller and $Y(2)$ will drift back to its original level. To illustrate a situation in which a CLMDM would be a useful model, consider an ice-cream market. Suppose that:

$$Y_t(1) = \log(\text{total sales of ice-cream})$$

$$Y_t(2) = \log(\text{market share of ice-cream type A})$$

Now if there was a heat wave, the total demand $Y_t(1)$ of ice-cream might suddenly increase. If brand A had extra stock and could cope with the extra demand better than another brand B, say, then the market share $Y_t(2)$ would be expected to increase over the heat wave. This situation could be modelled by a CLMDM such that:

$$Y_t(2) = \theta_t + \beta_t \{Y_t(1) - E[Y_t(1) | \mathbf{y}^{t-1}]\} + \epsilon_t,$$

where ϵ_t is some error term and β_t is some parameter associated with A's ability to cope with extra demand.

For both the LMDM and the CLMDM, the separate components $\{Y_t(r) | \mathbf{x}_t(r)\}$ have either a Gaussian or Student-t distribution depending on whether the forecast distribution is assumed known or estimated. The forecast distributions and updating relationships associated with these conditional univariate models are identical to those introduced in sections 3.1 and 3.3 respectively. It then follows

from theorem 5.4.1, that the one-step ahead forecast density of Y_t is simply the product of the univariate conditional one-step ahead forecast densities of $\{Y_t(r) | \mathbf{x}_t(r)\}$, for $r = 1, \dots, n$. Although the joint forecast distribution of \mathbf{y}_t will not, in general, be Gaussian, its mean and covariance matrix take a relatively simple form and these will now be derived here. Assume throughout this derivation that all means and variances are found conditionally on $\tilde{\mathbf{x}}_t(r)$.

The mean of the marginal forecast distribution for Y_t when it follows a LMDM will firstly be derived. As has already been mentioned, each variable under the LMDM follows a univariate regression DLM. Therefore from equation 3.6 the mean of the conditional forecast distribution for $\{Y_t(r) | \mathbf{x}_t(r)\}$ given the past is given by:

$$E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r)\} = F_t(r)^T \mathbf{a}_t(r) \quad (5.19)$$

where $\mathbf{a}_t(r)$ is an s_r -dimensional vector. Now suppose that $\mathbf{a}_t(r)$ is partitioned so that $\mathbf{a}_t(r)^T = (\mathbf{a}_t^*(r)^T, \tilde{\mathbf{a}}_t(r)^T)$ where $\mathbf{a}_t^*(r)^T = (a_t^{(1)}(r), \dots, a_t^{(r-1)}(r))$ contains those parameters associated with $\mathbf{x}_t(r)$ and $\tilde{\mathbf{a}}_t(r)$ is an $(s_r - r + 1)$ -dimensional vector containing those parameters associated with $\{\tilde{\mathbf{x}}_t(r), \mathbf{y}^{t-1}(r), \mathbf{x}^{t-1}(r)\}$. The expectation of the conditional forecast distribution of $\{Y_t(r) | \mathbf{x}_t(r)\}$ given the past can then be rewritten as:

$$\left. \begin{aligned} E \{Y_t(1) | \mathbf{y}^{t-1}(1)\} &= a_t^{(0)}(1) \\ E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r)\} &= a_t^{(0)}(r) + \sum_{i=1}^{r-1} a_t^{(i)}(r) Y_t(i) \end{aligned} \right\} \quad 2 \leq r \leq n \quad (5.20)$$

where $a_t^{(0)}(r)$ is a function of $\{\tilde{\mathbf{a}}_t(r), \tilde{\mathbf{x}}_t(r), \mathbf{y}^{t-1}(r), \mathbf{x}^{t-1}(r)\}$ only.

The marginal forecast means for Y_t given the past can then be easily calculated from these conditional means using the identity:

$$E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^{t-1}(r)\} = E \left[E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r)\} \right] \quad (5.21)$$

Therefore, from equation 5.20

$$E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^{t-1}(r)\} = a_t^{(0)}(r) + \sum_{i=1}^{r-1} a_t^{(i)}(r) E \{Y_t(i) | \mathbf{y}^{t-1}(i), \mathbf{x}^{t-1}(i)\}. \quad (5.22)$$

Write $\mathbf{a}_t^{(0)T} = (a_t^{(0)}(1), \dots, a_t^{(0)}(n))$ and A_t as the $n \times n$ lower triangular matrix whose $(j, k)^{th}$ element a_{jk} is given by:

$$a_{jk} = \begin{cases} a_t^{(k)}(j) & k < j \\ 0 & \text{otherwise} \end{cases}$$

Then equation 5.22 across all $r = 1, \dots, n$ can be expressed by:

$$E(Y_t | \mathbf{y}^{t-1}) = \mathbf{a}_t^{(0)} + A_t E(Y_t | \mathbf{y}^{t-1})$$

and so

$$E(Y_t | \mathbf{y}^{t-1}) = [I - A_t]^{-1} \mathbf{a}_t^{(0)}.$$

Similarly for the CLMDM, the expectation of the forecast distribution for $\{Y_t(r) | \mathbf{x}_t(r)\}$ given the past can be rewritten as:

$$\left. \begin{aligned} E \{Y_t(1) | \mathbf{y}^{t-1}(1)\} &= \bar{a}_t^{(0)}(1) \\ E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r)\} &= \bar{a}_t^{(0)}(r) + \sum_{i=1}^{r-1} \bar{a}_t^{(i)}(r) \{Y_t(i) - \hat{f}_t(i)\} \end{aligned} \right\} \quad 2 \leq r \leq n \quad (5.23)$$

where $\bar{a}_t^{(j)}(r)$ corresponds to a different value of the analogous $a_t^{(j)}(r)$ of equation 5.20 and $\hat{f}_t(r)$ is the same as in equation 5.18. By using the identity 5.21 it is obvious that the marginal forecast means for $Y_t(r)$ following the CLMDM are simply:

$$E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^{t-1}(r)\} = \bar{a}_t^{(0)}(r).$$

Although the marginal covariance matrix of \mathbf{Y}_t given the past is found from the same recursive relationships in both the LMDM and the CLMDM, the covariance matrix of the CLMDM takes a simpler form and so only the derivation

of this matrix will be shown here. To find this matrix it is first necessary to find the variance of the marginal forecast distribution of $Y_t(r)$, for $r = 1, \dots, n$.

Suppose that $\theta_t(r)^T = \{\theta_t^*(r)^T, \tilde{\theta}_t(r)^T\}$ where $\theta_t^*(r)$ is the set of parameters contained in $\theta_t(r)$ associated with $\mathbf{x}_t(r)$ and $\tilde{\theta}_t(r)$ is the set associated with $\tilde{\mathbf{x}}_t(r)$. If $R_t(r)$ is the covariance matrix of the prior distribution of $\{\theta_t(r) | \mathbf{y}^{t-1}\}$ then $R_t(r)$ can be expressed by:

$$R_t(r) = \begin{pmatrix} R_t^*(r) & R_t'(r) \\ R_t'(r)^T & \tilde{R}_t(r) \end{pmatrix}$$

where

$$R_t^*(r) = \text{cov} \{ \theta_t^*(r), \theta_t^*(r) | \mathbf{y}^{t-1} \}, \quad \tilde{R}_t(r) = \text{cov} \{ \tilde{\theta}_t(r), \tilde{\theta}_t(r) | \mathbf{y}^{t-1} \}$$

$$\text{and } R_t'(r) = \text{cov} \{ \theta_t^*(r), \tilde{\theta}_t(r) | \mathbf{y}^{t-1} \}$$

Therefore, from equation 3.6, the variance of the conditional forecast distribution of $\{Y_t(r) | \mathbf{x}_t(r)\}$ given the past is given by:

$$\text{var} \{ Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r) \} = \mathbf{F}_t(r)^T R_t(r) \mathbf{F}_t(r) + V_t(r).$$

Let $\tau_t^2(r) \geq 0$, $r = 1, \dots, n$ be a function of the known constants $\tilde{\mathbf{x}}_t(r)$, $\mathbf{y}^{t-1}(r)$, $\mathbf{x}^{t-1}(r)$, $\tilde{R}_t(r)$, $R_t'(r)$ and $V_t(r)$. The variance of the marginal forecast distribution of Y_t following a CLMDM then becomes:

$$\left. \begin{aligned} \text{var} \{ Y_t(1) | \mathbf{y}^{t-1}(1) \} &= \tau_t^2(1) \\ \text{var} \{ Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r) \} &= \tau_t^2(r) + \{ \mathbf{x}_t(r) - \hat{\mathbf{f}}_t(r) \}^T R_t^*(r) \{ \mathbf{x}_t(r) - \hat{\mathbf{f}}_t(r) \} \end{aligned} \right\} \quad 2 \leq r \leq n$$

Now, since $R_t^*(r)$ is a covariance matrix it is positive definite and so can be written in the form:

$$R_t^*(r) = S_t(r) S_t(r)^T$$

where $S_t(r)$ is a non-singular matrix. Therefore the variance can be rewritten as:

$$\begin{aligned} \text{var} \{ Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r) \} &= \tau_t^2(r) + \{ \mathbf{x}_t(r) - \hat{\mathbf{f}}_t(r) \}^T S_t(r) S_t(r)^T \{ \mathbf{x}_t(r) - \hat{\mathbf{f}}_t(r) \}, \\ & \quad 2 \leq r \leq n. \end{aligned}$$

Using the property that:

$$\text{trace}(AB) = \text{trace}(BA)$$

and the fact that the trace of a scalar is merely that scalar, the variance becomes:

$$\left. \begin{aligned} \text{var} \{Y_t(1) | \mathbf{y}^{t-1}(1)\} &= \tau_t^2(1) \\ \text{var} \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r)\} &= \tau_t^2(r) + h(\mathbf{x}_t(r)), \quad 2 \leq r \leq n \end{aligned} \right\} \quad (5.24)$$

where

$$h(\mathbf{x}_t(r)) = \tau_t^2(r) + \text{trace} \left[S_t(r)^T \{ \mathbf{x}_t(r) - \hat{\mathbf{f}}_t(r) \} \{ \mathbf{x}_t(r) - \hat{\mathbf{f}}_t(r) \}^T S_t(r) \right].$$

To find the variance of the marginal forecast distribution of $Y_t(r)$ given the past, it is first noted that for any two variables $Z(1)$ and $Z(2)$, the following identity holds:

$$\text{var} \{Z(2)\} = E [\text{var} \{Z(2) | Z(1)\}] + \text{var} [E \{Z(2) | Z(1)\}]. \quad (5.25)$$

Suppose that Σ_t is the forecast covariance matrix for Y_t such that:

$$\{\Sigma_t\}_{jk} = \sigma_t(j, k) = \text{cov} \{Y_t(j), Y_t(k) | \mathbf{y}^{t-1}\}, \quad j, k = 1, \dots, n$$

and let $\Sigma_t(r)$ be the forecast covariance matrix of $\{Y_t(1), \dots, Y_t(r-1)\}$ for $r = 2, \dots, n$. Therefore by using identity 5.25 together with equations 5.23 and 5.24, it can be shown that:

$$\left. \begin{aligned} \sigma_t(1, 1) &= \tau_t^2(1) \\ \sigma_t(r, r) &= E [\text{var} \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r)\}] + \text{var} [E \{Y_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^t(r)\}] \\ &= \tau_t^2(r) + \text{trace} \{ S_t(r)^T \Sigma_t(r) S_t(r) \} + \bar{\mathbf{a}}_t(r)^T \Sigma_t(r) \bar{\mathbf{a}}_t(r) \end{aligned} \right\} \quad 2 \leq r \leq n$$

where $\bar{\mathbf{a}}_t(r) = (\bar{a}_t^{(1)}(r), \dots, \bar{a}_t^{(r-1)}(r))$.

To find the marginal covariance between $Y_t(r)$ and $Y_t(k)$, note that for two variables $Z(1)$ and $Z(2)$

$$E \{Z(1)Z(2)\} = E [E \{Z(2)Z(1) | Z(2)\}] = E [Z(2) \cdot E \{Z(1) | Z(2)\}]$$

Therefore suppose that $Z(1) = Y_t(r)$ and $Z(2) = X_t(r)$ and let $\sigma_t(r)$ be a vector such that for $r \geq 2$

$$\sigma_t(r)^T = (\sigma_t(r, 1), \dots, \sigma_t(r, r-1)) = \text{cov}\{Y_t(r), X_t(r) | y^{t-1}\}.$$

Then

$$\sigma_t(r) = E \left[(X_t(r), E\{Y_t(r) | X_t(r)\}) | y^{t-1} \right].$$

By equation 5.23 this becomes:

$$\begin{aligned} &= E \left[X_t(r) \left(\bar{a}_t^{(0)}(r) + \sum_{i=1}^{r-1} \bar{a}_t^{(i)}(r) \{Y_t(i) - \hat{f}_t(i)\} \right) | y^{t-1} \right] \\ &= \bar{a}_t^{(0)}(r) E \{X_t(r) | y^{t-1}\} + \sum_{i=1}^{r-1} \bar{a}_t^{(i)}(r) E [X_t(r) \{Y_t(i) - \hat{f}_t(i)\} | y^{t-1}] \end{aligned}$$

It is clear that:

$$\Sigma_t(r+1) = \begin{pmatrix} \Sigma_t(r) & \sigma_t(r) \\ \sigma_t(r)^T & \sigma_t(r, r) \end{pmatrix}$$

So if $\Sigma_t(2) = \sigma(1, 1)$ is given, then $\Sigma_t(r+1)$ can be calculated as a simple function of $\Sigma_t(r)$, $\sigma_t(r)$ and $\sigma_t(r, r)$. Hence the marginal forecast covariance matrix of the CLMDM, is simply calculated from the variances and expectations of the updated distributions for the separate conditional component regression DLM's.

These models are best illustrated by some simple examples.

5.5.1 Example.

The simplest nontrivial example which illustrates this new class of models can be constructed as follows. Suppose that $Y_t^T = (Y_t(1), Y_t(2))$ is to be modelled by a LMDM and that at any fixed time the two variables can be represented by the graph of the influence diagram in figure 5.8(a). Suppose further that the processes $\{Y_k(1)\}_{k \leq t}$ and $\{Y_k(2)\}_{k \leq t}$ can be represented by the graph of the influence diagram of figure 5.8(b).

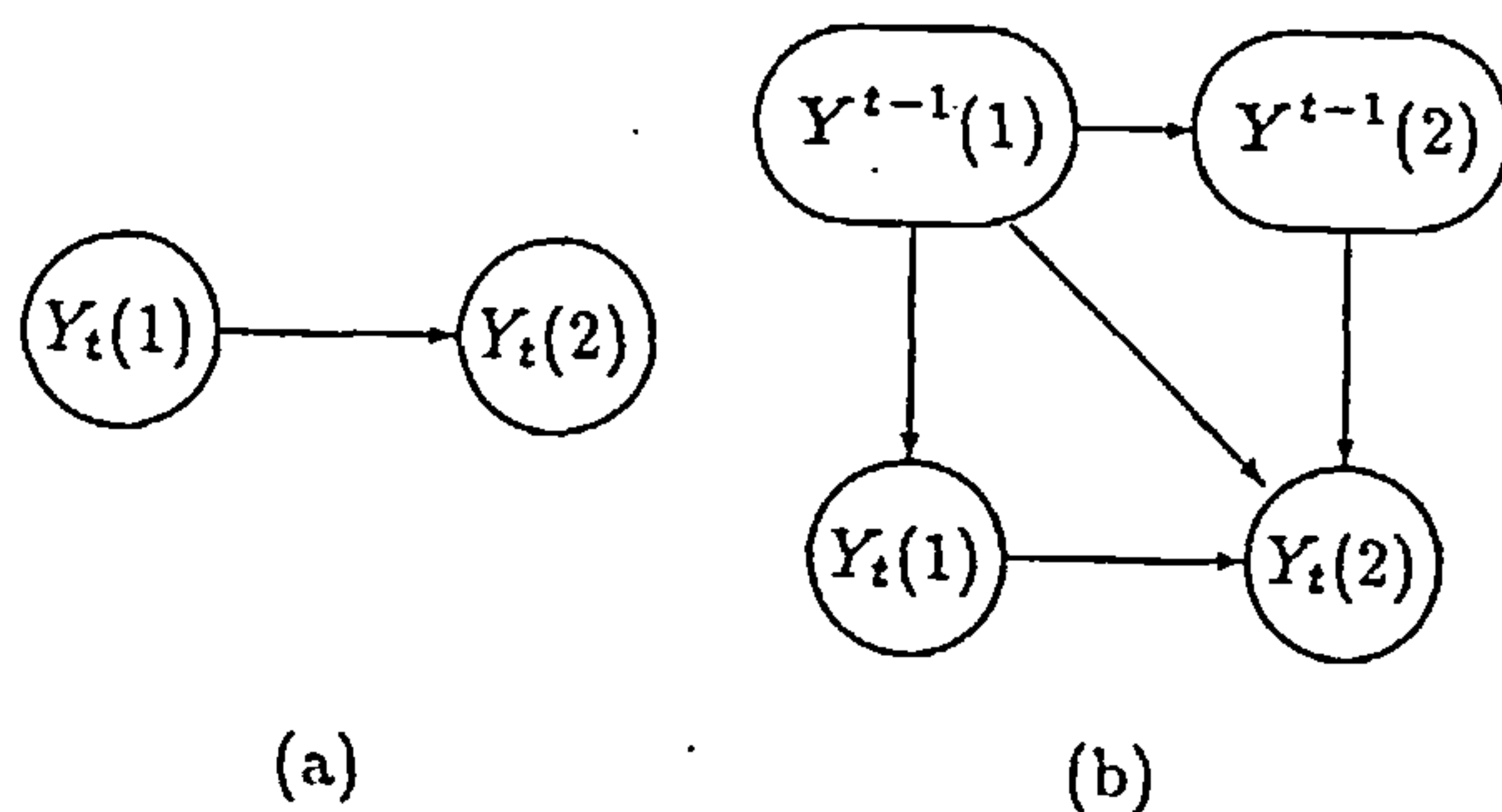


Figure 5.8: Graph of the influence diagrams for (a) $Y(1)$ and $Y(2)$ at a fixed time t and (b) $\{Y_k(1)\}_{k \leq t}$ and $\{Y_k(2)\}_{k \leq t}$.

Let the following observation and system equations hold:

Observation Equations

$$Y_t(1) = \theta_t(1) + v_t(1), \quad v_t(1) \sim N(0, V_t(1))$$

$$Y_t(2) = y_t(1)\theta_t(2) + v_t(2), \quad v_t(2) \sim N(0, V_t(2))$$

System Equation

$$\theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim N(0, W_t)$$

Initial information

$$(\theta_0 | D_0) \sim N(m_0, C_0)$$

where G_t is the 2×2 identity matrix; W_t and C_0 are diagonal; and $v_t(1)$, $v_t(2)$, $w_t(1)$ and $w_t(2)$ are mutually independent of each other across time. Notice that in this example $F_t(1) = 1$ and $F_t(2) = y_t(1)$.

Suppose that at time $t - 1$ the information about the parameters is expressed through the distribution:

$$(\theta_{t-1} | y^{t-1}) \sim N(m_{t-1}, C_{t-1})$$

where $m_{t-1}^T = (m_{t-1}(1), m_{t-1}(2))$ and C_{t-1} is diagonal.

From equation 3.6 and corollary 5.4.2, it can be immediately seen that the following means and variances for the conditional forecast distributions come through:

$$\begin{aligned} E\{Y_t(1) | \mathbf{y}^{t-1}(1)\} &= F_t(1)m_{t-1}(1) = m_{t-1}(1) \\ E\{Y_t(2) | \mathbf{y}^{t-1}, y_t(1)\} &= F_t(2)m_{t-1}(2) = y_t(1)m_{t-1}(2) \end{aligned}$$

and

$$\begin{aligned} \text{var}\{Y_t(1) | \mathbf{y}^{t-1}(1)\} &= \tau_t^2(1) \\ \text{var}\{Y_t(2) | \mathbf{y}^{t-1}, y_t(1)\} &= y_t(1)^2 R_t^*(2) + \tau_t^2(2) \end{aligned}$$

where $\tau_t^2(1)$ just depends on t and follows the usual DLM iterative equations, $\tau_t^2(2)$ is simply $V_t(2)$ and $\text{trace}[S_t(2)^T \{x_t(2)\} \{x_t(2)\}^T S_t(2)] = y_t(1)^2 R_t^*(2)$, where $R_t^*(2)$ (which is equivalent to $R_t(2)$ using the notation of sections 3.1 and 5.5) is a function of $\mathbf{y}^{t-1}(1)$.

Thus, by using the results introduced earlier in this section, it is clear that the forecast distribution of Y_t will have mean vector:

$$\begin{aligned} E[Y_t | \mathbf{y}^{t-1}] &= \left[I - \begin{pmatrix} 0 & 0 \\ m_{t-1}(2) & 0 \end{pmatrix} \right]^{-1} \begin{pmatrix} m_{t-1}(1) \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} m_{t-1}(1) \\ m_{t-1}(1)m_{t-1}(2) \end{pmatrix} \end{aligned}$$

and covariance matrix:

$$\Sigma = \begin{pmatrix} \tau_t^2(1) & \sigma_t(1, 2) \\ \sigma_t(2, 1) & \sigma_t(2, 2) \end{pmatrix}$$

where

$$\begin{aligned} \sigma_t(1, 2) &= \sigma_t(2, 1) = m_{t-1}(2) E\{Y_t(1)^2 | \mathbf{y}^{t-1}\}, \\ \sigma_t(2, 2) &= \tau_t^2(2) + R_t^*(2) E\{Y_t(1)^2 | \mathbf{y}^{t-1}\} + m_{t-1}(2)^2 \tau_t^2(1). \end{aligned}$$

Clearly, both $\{Y_t(1)|\mathbf{y}^{t-1}(1)\}$ and $\{Y_t(2)|\mathbf{y}^{t-1}, y_t(1)\}$ have normal distributions. However, the joint forecast distribution of $\{Y_t(1), Y_t(2)|\mathbf{y}^{t-1}\}$ is *not* bivariate normal because of the appearance of $y_t(1)$ in the variance term of $\{Y_t(2)|\mathbf{y}^{t-1}, y_t(1)\}$. Indeed this joint distribution can be very non-normal. This is demonstrated in figures 5.9 and 5.10 in which the contours of the joint distribution of $\{Y_t(1), Y_t(2)\}$ after the margin of $Y_t(1)$ has been normalised and for various different parameter values are given.

From these diagrams it can be shown that the modes and anti-modes lie on a quintic (and so exhibit a Butterfly catastrophe, Zeeman, 1977). Notice that the joint forecast distribution is only symmetrical when the variance of $\{Y_t(1)|\mathbf{y}^{t-1}\}$ does not appear in the variance of $\{Y_t(2)|\mathbf{y}^{t-1}\}$, that is, when $m_{t-1}(2) = 0$. For any non-degenerate values of the parameters, it can be concluded, after a little algebra, that there is a value of $y_t(2)$ such that the conditional predictive density of $\{Y_t(1)|Y_t(2) = y_t(2)\}$ is bimodal. The worst case obviously occurs when $\theta_t(2)$ is uncertain where the distribution becomes very non-Gaussian. As $R_t^*(2) \rightarrow 0$, then the process tends to a bivariate normal. However, unlike the analogous simultaneous equations models the assumed stochastic drift on $\theta_t(2)$ prevents this limit from ever being reached. 3-D plots of the same LMDM when the observation variances are unknown are shown in figure 5.11. Notice how the distributions are now also dependent on the number of degrees of freedom in the t-distribution of each variable.

From the figures it is clear that the point $E[Y_t(1)|\mathbf{y}^{t-1}(1)]$ (in this case $Y_t(1) = 0$) takes on special significance as it is the point about which the contours are symmetrical or asymmetrical. Regression on the forecast residual $[Y_t(1) - E\{Y_t(1)|\mathbf{y}^{t-1}(1)\}]$ will often therefore seem more natural. This gives

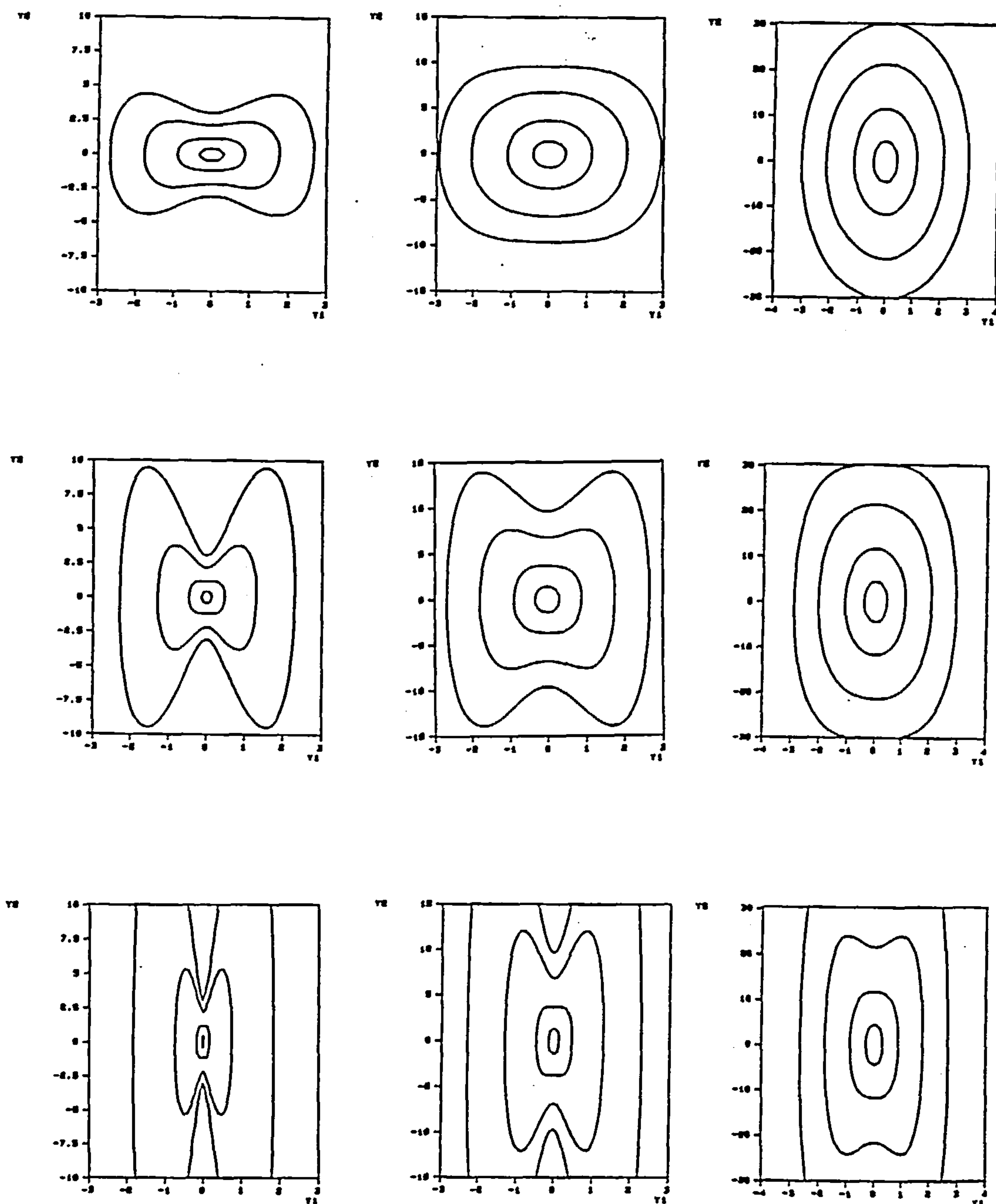


Figure 5.9: Various symmetric contour plots of $Y_t(1)$ and $Y_t(2)$ when the marginal distribution of $Y_t(1)$ has been normalised to have zero mean and unit variance and $m_{t-1}(2) = 0$. $R_t^*(2)$ is kept constant for all graphs in the same row and $\tau_t^2(2)$ is kept constant for each column. The value of both $R_t^*(2)$ and $\tau_t^2(2)$ in the first row/column is 1, their value in the second row/column is 10 and in the third is 100.

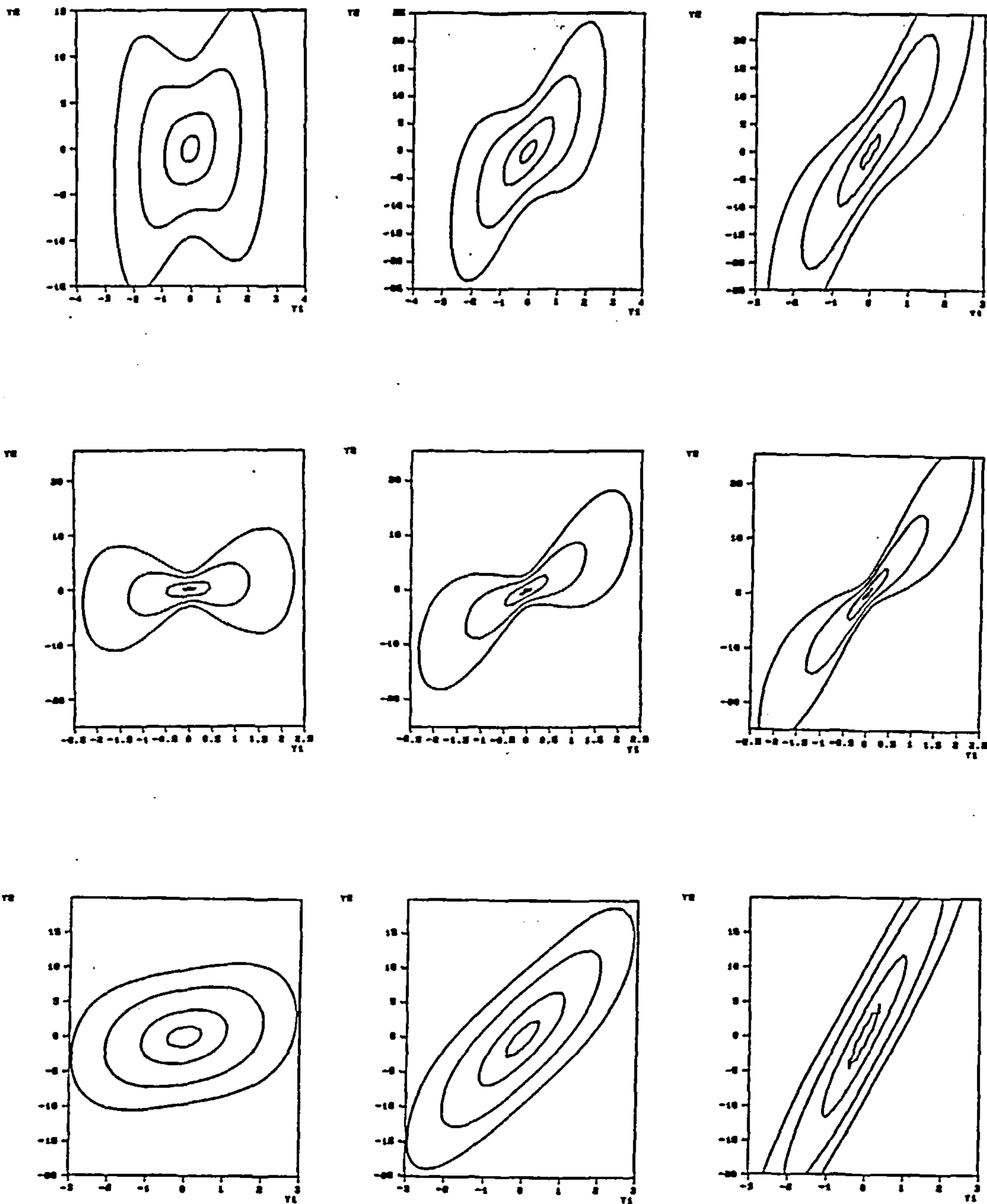


Figure 5.10: Various asymmetric contour plots of $Y_i(1)$ and $Y_i(2)$ when the marginal distribution of $Y_i(1)$ has been normalised to have zero mean and unit variance and $m_{i-1}(2) \neq 0$. $R_i^*(2) = \tau_i^2(2) = 10$ in the first row, $R_i^*(2) = 10$, $\tau_i^2(2) = 1$ in the second row and $R_i^*(2) = 1$, $\tau_i^2(2) = 10$ in the third. $m_{i-1}(2)$ varies by column, taking the values 1 in the first column, 5 in the second and 10 in the third.

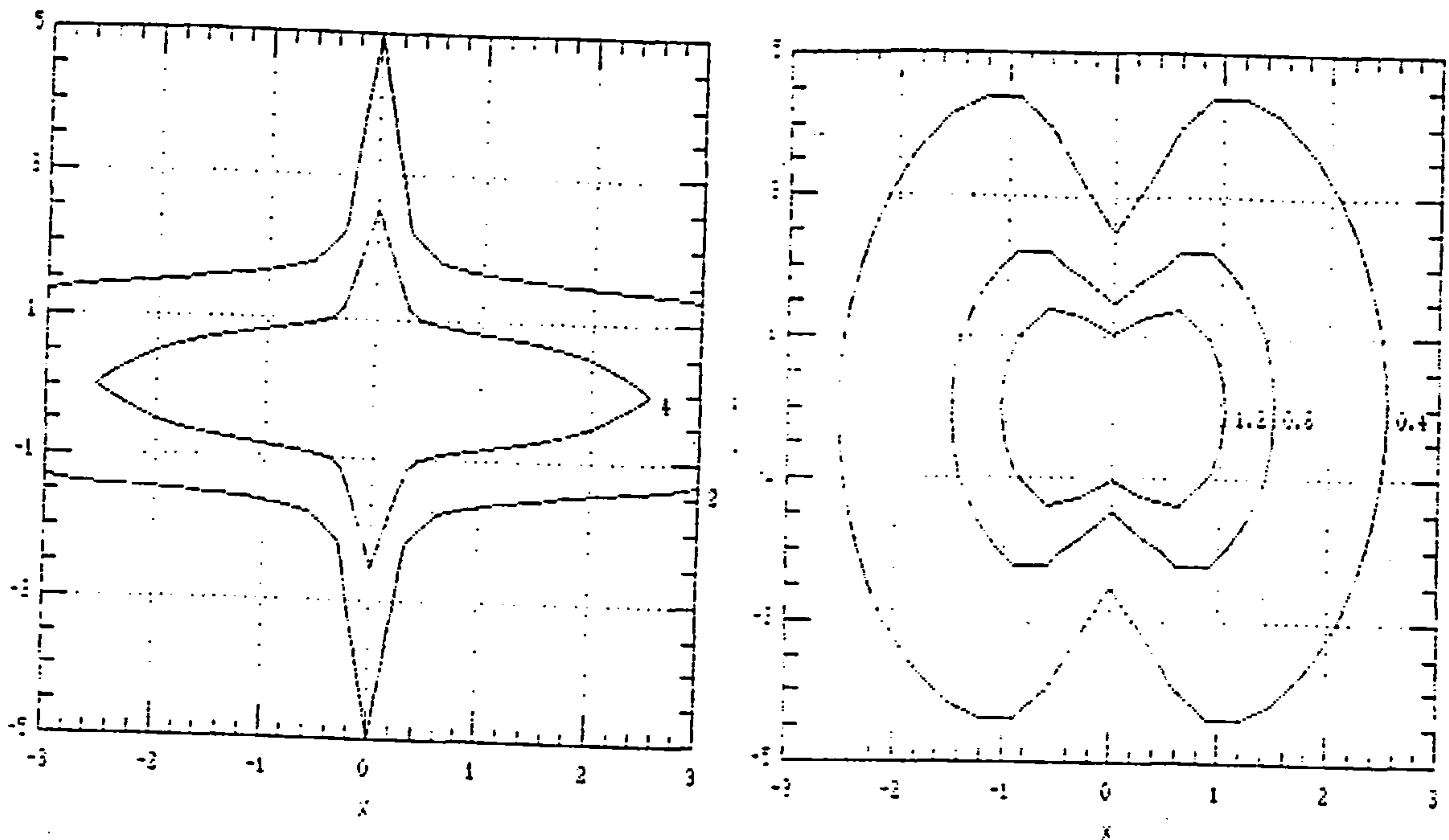


Figure 5.11: 3-D plots of joint t-distribution of 2 variables following an LMDM. Notice how the contour becomes star shape as the degrees of freedom get smaller.

the corresponding CLMDM whose means and covariance matrix are given by:

$$E[Y_t | \mathbf{y}^{t-1}] = \begin{pmatrix} m_{t-1}(1) \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \tau_t^2(1) & \sigma_t(1, 2) \\ \sigma_t(2, 1) & \sigma_t(2, 2) \end{pmatrix}$$

where

$$\sigma_t(1, 2) = \sigma_t(2, 1) = m_{t-1}(2) E \left[\{Y_t(1) - \hat{f}_t(1)\}^2 | \mathbf{y}^{t-1} \right],$$

$$\begin{aligned} \sigma_t(2, 2) &= \tau_t^2(2) + R_t^*(2) E \left[\{Y_t(1) - \hat{f}_t(1)\}^2 | \mathbf{y}^{t-1} \right] \\ &\quad + m_{t-1}(2)^2 \text{var} \left[\{Y_t(1) - \hat{f}_t(1)\} | \mathbf{y}^{t-1} \right]. \\ &= \tau_t^2(2) + \tau_t^2(1) \{R_t^*(2) + m_{t-1}(2)^2\}. \end{aligned}$$

The graphs in figures 5.9 and 5.10 will still be appropriate for this corrected model.

5.5.2 Example.

Consider now the more general situation for the 2 variable case, where G_t is no longer the identity, $F_t(1) = \tilde{x}_t(1)$ and $F_t(2)^T = (y_t(1), \tilde{x}_t(2))$. The general LMDM now holds:

Observation Equation

$$\begin{aligned} Y_t(1) &= F_t(1)^T \theta_t(1) + v_t(1), & v_t(1) &\sim N(0, V_t(1)) \\ Y_t(2) &= F_t(2)^T \theta_t(2) + v_t(2), & v_t(2) &\sim N(0, V_t(2)) \end{aligned}$$

System Equation

$$\theta_t = G_t \theta_{t-1} + w_t, \quad w_t \sim N(\mathbf{0}, W_t)$$

Initial information

$$(\theta_0 | D_0) \sim N(m_0, C_0)$$

where G_t , W_t and C_0 are blockdiagonal as usual and the same independence conditions hold as for example 5.5.1

Using exactly the same notation introduced to define the means and variances of the LMDM in this section, equations 5.20 and 5.24 lead directly to the required conditional forecast means and variances which are given by:

$$\begin{aligned} E\{Y_t(1) | y^{t-1}(1)\} &= a_t^{(0)}(1) \\ E\{Y_t(2) | y^{t-1}, y_t(1)\} &= a_t^{(0)}(2) + y_t(1)a_t^{(1)}(2) \end{aligned}$$

and:

$$\begin{aligned} \text{var}\{Y_t(1) | y^{t-1}(1)\} &= \tau_t^2(1) \\ \text{var}\{Y_t(2) | y^{t-1}, y_t(1)\} &= y_t(1)^2 R_t^*(2) + \tau_t^2(2) \end{aligned}$$

The only qualitative difference from Example 5.5.1 is that the iterative form of $\tau_t^2(2)$ now depends on $y^{t-1}(1)$. This is because $\tau_t^2(2)$ is now a function of

$\tilde{x}_t(2)$, $\tilde{R}_t(2)$, $R'_t(2)$ and $V_t(2)$ where $\tilde{R}_t(2)$ and $R'_t(2)$ will depend on $\mathbf{y}^{t-1}(1)$. In particular, the first two moments of the forecast distribution are simply:

$$E[Y_t | \mathbf{y}^{t-1}] = \begin{pmatrix} a_t^{(0)}(1) \\ a_t^{(0)}(2) + a_t^{(0)}(1)a_t^{(1)}(2) \end{pmatrix}$$

and covariance matrix:

$$\Sigma = \begin{pmatrix} \tau_t^2(1) & \sigma_t(1, 2) \\ \sigma_t(2, 1) & \sigma_t(2, 2) \end{pmatrix}$$

where

$$\sigma_t(1, 2) = \sigma_t(2, 1) = a_t^{(0)}(2)E\{Y_t(1) | \mathbf{y}^{t-1}\} + a_t^{(1)}(2)E\{Y_t(1)^2 | \mathbf{y}^{t-1}\}$$

and

$$\sigma_t(2, 2) = \tau_t^2(2) + R_t^*(2)E\{Y_t(1)^2 | \mathbf{y}^{t-1}\} + a_t^{(1)}(2)^2\tau_t^2(1).$$

It can be seen that the geometry of the problem, given this more complicated model, just parallels that of example 5.5.1.

5.6 Discussion of the MDM.

Not only do MDM's accommodate certain features associated with partial segmentation and causal structures of business multivariate time series, but they are also of theoretical interest. Some of the theoretical aspects of MDM's will be discussed in this section.

MDM's define a class of non-Gaussian time series models which decompose the forecasting system into components whose conditional distributions are univariate Bayesian dynamic regression models. As was mentioned in sections 5.3 and 5.5, these univariate models can be normal. In particular, if $F_t(r)$ is a linear function of $\{\mathbf{x}^t(r), \mathbf{y}^{t-1}(r)\}$, then each variable would follow a generalisation of one of

Priestley's state-dependent models (Priestley, 1980). Alternatively, each model could be a non-linear Dynamic Model (West, Harrison & Migon, 1985, Migon & Harrison, 1985, Pole, West & Harrison, 1988) enabling any non-linear effects of competitive strategies to be modelled directly. If each component follows a univariate Bayesian linear or non-linear model, then it is possible to use the updating relationships directly from univariate Bayesian dynamic models, even though the model regresses on contemporary components of Y_t and can be highly non-linear. This makes the process especially interesting because the models are amenable to analytical investigation. Approximate or numerical methods, with all the robustness issues that surround them, are largely unnecessary. On the other hand, as can be seen from figures 5.9 and 5.10 of example 5.5.1, the vector Y_t can have a joint forecast distribution which is very non-Gaussian, even when F_t is a linear function of $(Y_t(1), \dots, Y_t(r-1))$. In this respect this class of models is very similar to the models of Wermuth & Lauritzen (1990) in the fact that the conditional distributions are fairly simple but the joint model is far more complex.

Since the forecasting model is expressed in terms of univariate Bayesian dynamic models, such techniques as intervention, trends and seasonals, can be transferred directly on to these models. Notice that the MDM does not impose the stringent symmetry conditions necessary in the models of Harvey, 1986 and the DMR models (see section 3.6), that is, the MDM does not require that each variable has the same F_t and G_t in its observation and system equations. The forecaster need not specify fixed values for the observation variances $V_t(1), \dots, V_t(n)$ and the variance for each variable can be estimated on-line with the system as described in section 3.3. Discount factors can also be used to specify values for the system errors $w_t(1), \dots, w_t(n)$ (see section 3.2).

It is interesting to note that, unlike conditional independence within time

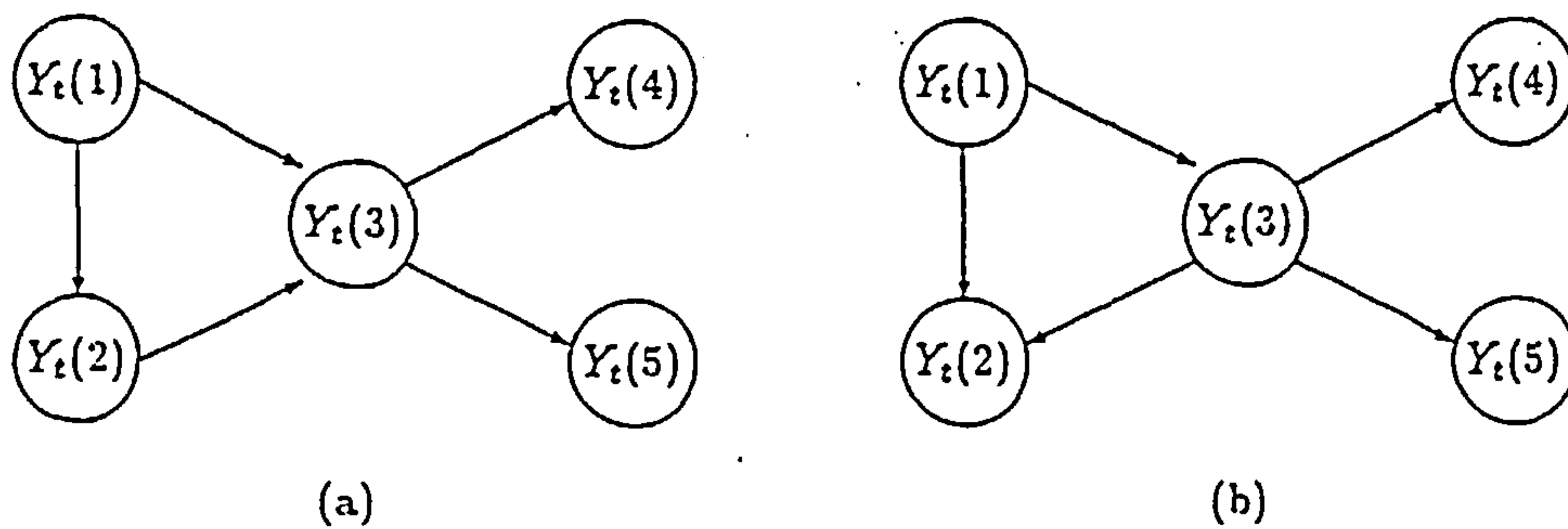


Figure 5.12: Two graph with the same implied conditional independences, but with different LMDM's.

frames, conditional causality of the type described above depends on the ordering of the variables in the MDM. Furthermore each influence diagram corresponds to a unique LMDM or CLMDM structure.

For example, suppose that for 5 variables modelled with an LMDM, the following regression equations hold:

$$\left. \begin{aligned} Y_t(1) &= \theta_t^{(0)}(1) + v_t(1) \\ Y_t(2) &= \theta_t^{(0)}(2) + \theta_t^{(1)}(2)y_t(1) + v_t(2) \\ Y_t(3) &= \theta_t^{(0)}(3) + \theta_t^{(1)}(3)y_t(1) + \theta_t^{(2)}(3)y_t(2) + v_t(3) \\ Y_t(4) &= \theta_t^{(0)}(4) + \theta_t^{(3)}(4)y_t(3) + v_t(4) \\ Y_t(5) &= \theta_t^{(0)}(5) + \theta_t^{(3)}(5)y_t(3) + v_t(5) \end{aligned} \right\} \quad (5.26)$$

The graph of the influence diagram consistent with these equations is given by graph (a) in figure 5.12.

It is easy to check that the LMDM on these regression equations, could alternatively be represented by the LMDM on components of Y taken in the order $\{Y(1), Y(2), Y(3), Y(5), Y(4)\}$. The graph of the influence diagram figure 5.12(a) would remain unchanged.

In general any new ordering of the variables compatible with the influence diagram of a single time frame of a LMDM or CLMDM, given the past, will

produce equations algebraically identical to the original.

On the other hand, note that by the Decomposition Theorem (Smith, 1989), for any given time frame t , the implied set of conditional independence statements by graph (b) in figure 5.12, given the past, are identical to those represented by graph (a). However, the conditional independences *related to causality* in the two diagrams, correspond to two quite different LMDM's.

For example, from graph (a) it is clear that:

$$Y_t(2) | \mathbf{y}^{t-1} = Y_t(2) | \mathbf{y}^{t-1}(2), \mathbf{y}^{t-1}(1)$$

whereas graph (b) means that:

$$Y_t(2) | \mathbf{y}^{t-1} = Y_t(2) | \mathbf{y}^{t-1}(2), \mathbf{y}^{t-1}(1), \mathbf{y}^{t-1}(3).$$

Because of the different covariance structure on the system error w_t implied by each of these influence diagrams, there is no guarantee that these two statements could hold simultaneously.

Suppose that the context of the model is such that the time frame conditional independences are logically determined. However, suppose that there is uncertainty regarding the causal structure across the variables in the problem, although this causal structure is assumed consistent over time. In this case a Class I *Multi-process MDM* can model the system. A multi-process MDM has m models $\{M^{(1)}, \dots, M^{(m)}\}$ where there is one model for each of the m possible causal structures for the given conditional independence structure. The methodology outlined for univariate multi-process models in section 3.5 can then be applied directly. Multi-process MDM's allow the prediction of complex series without any apriori assumptions about causal structures. This is because the forecasts are found by using $p(\mathbf{y}_t | \mathbf{y}^{t-1}, M^{(i)})$ mixed with probabilities $p(M^{(i)} | \mathbf{y}^{t-1})$ to give appropriate predictive densities.

Although Granger causality has attracted academic interest, in practice it has proved difficult to discriminate between different causal structures using linear systems. However, by embedding causality in the non-linear MDM, multi-process MDM's can give an *on-line* assessment of hypothetical causal relationships across the variables. The MDM corresponding to a given causal structure can then be made at least plausible. Furthermore, the conditional components can be as complicated as necessary, containing trends, regression terms, seasonal factors, and so on. It therefore looks promising that the MDM's will enable the selection of causal structures across practical models.

Notice that from the geometries of the joint densities on the model of examples 5.5.1 and 5.5.2, most information about causal relationships seems to come when the relationship between the variables is uncertain (in these examples this corresponds to $R_i^*(2)$ being large). This will occur early in the series and after external intervention (see section 3.4). This might explain why Zellner (1987) finds it so difficult to discriminate between two causal structures — the models he considers are not dynamic, they would assume that $R_i^*(2) \rightarrow 0$ and external intervention is not considered.

Chapter 6

Dynamic Graphical Models.

6.1 Introduction.

Chapter 5 introduced MDM's which are not only fairly simple to implement, but are also flexible enough to accommodate many of the causal relationships which can exist amongst brand sales in competitive business markets. MDM's only have a restricted use, however, as they assume that all components in a vector time series Y_t are causally linked and they do not allow any symmetries to exist amongst the components. *Dynamic graphical models* (DGM's) attempt to overcome this problem, modelling both the causal and symmetric relationships which might exist between components.

DGM's are a combination of (conditionally normal) MDM's and DMR models (see section 3.6). They attempt to use the flexibility of MDM's to model any causal relationships whilst utilising the imposed symmetry of the DMR model to accommodate any natural symmetries that might possibly exist amongst the components of a vector time series Y_t . They are essentially (conditionally normal) MDM's in which some of the components are vectors and the observation covariance matrix, which is assumed unknown, is estimated on-line with the system.

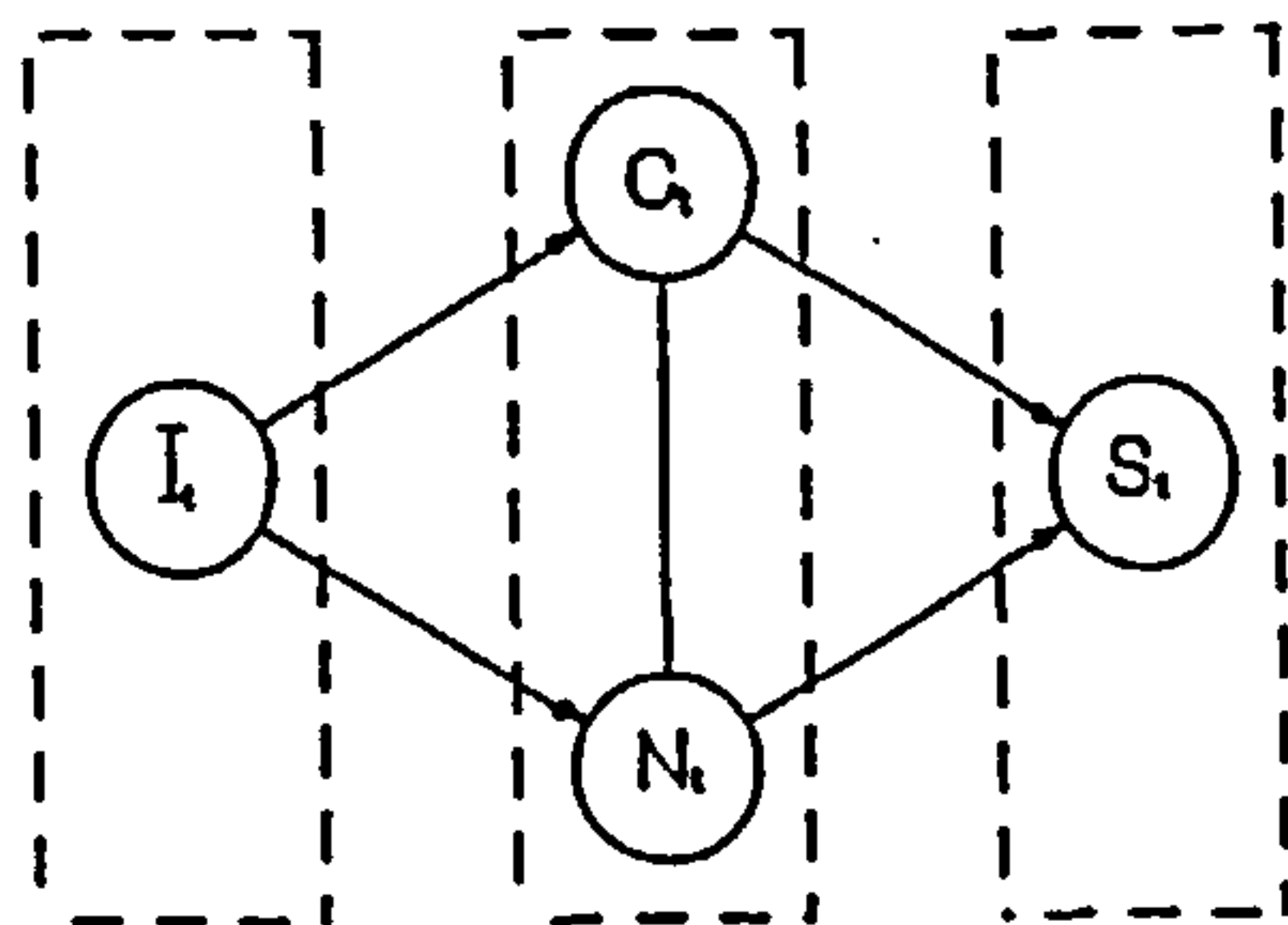


Figure 6.1: Graph representing the research hypothesis of the shampoo market.

To illustrate how a DGM is derived for a multivariate time series, consider the problem of forecasting sales in a hypothetical small shampoo market. Consumers can choose between buying equally high priced conditioning shampoo and natural ingredients shampoo, or a cheaper standard shampoo. Because the standard shampoo is not very attractive, consumers would prefer to buy one of the more expensive shampoos if they can afford them. So it is expected that, at time t , the sales of the more expensive conditioning and natural ingredients shampoos, C_t and N_t respectively, will decrease as the number of affluent customers decreases, the rest having to buy standard shampoo. An index, I_t , of disposable income amongst consumers, at time t , may be a good indicator of consumer affluence. Let the sales of standard shampoo at time t be represented by S_t , then a research hypothesis of the causal links between the series at a single time period can be seen in the graphical representation of figure 6.1.

If edge (C_t, N_t) of figure 6.1 were directed, then this would be the graph of an influence diagram. In this case, the results of chapter 5 would mean that $\{I_t\}_{t \leq 1}$ is a Granger non-cause of $\{S_t\}_{t \leq 1}$ given $\{C_t, N_t\}_{t \leq 1}$ when it is modelled by an MDM.

However, an MDM would require that N_t be conditioned on (C_t, I_t) , but C_t would be conditioned on I_t alone, thus destroying the symmetry of the role

between C_t and N_t in the given explanation of the dynamics of the market, as well as introducing spurious non-causes. It is therefore inappropriate to represent the series by an influence diagram. A DGM is a more satisfactory model for this situation as it would accommodate the symmetric relationship between $\{C_t\}$ and $\{N_t\}$ represented by the undirected edge (C_t, N_t) in figure 6.1 whilst $\{I_t\}_{t \leq 1}$ remains a conditional non-cause of $\{S_t\}_{t \geq 1}$ given $\{C_t, N_t\}_{t \geq 1}$. Thus both the causal and symmetric relationships of figure 6.1 are accommodated.

In terms of causal structures, these DGM's are represented by a subclass of chain graph (Wermuth & Lauritzen, 1990) at every time frame. Chain graphs are mixed graphs and they partition variables into subsets, which will be called blocks here, such that directed edges connect variables between blocks consistent with the order of the partition (lower indexed blocks to higher) and any pairs of variables within blocks are joined by a non-causal undirected edge. This subclass is defined by imposing the following two conditions:

1. *all* variables in the same block of the chain graph are adjacent
2. if there is a directed edge from a variable in one block to a variable in another block, then there must be a directed edge from *every* variable in the first block to *every* variable in the second.

Figure 6.1 represents a chain graph of causality that lies in the subclass defined by 1 and 2 above. The partition of blocks is $(\{I_t\}, \{C_t, N_t\}, \{S_t\})$. Figure 6.2 shows a research hypothesis of a real market containing 9 brands.

As for MDM's, the relationships between the present Y_t and the past series Y^{t-1} can be represented by a graph in which each variable in the chain graph at time t has as its parent set its own past series, as well as the past series of its parents at time t .

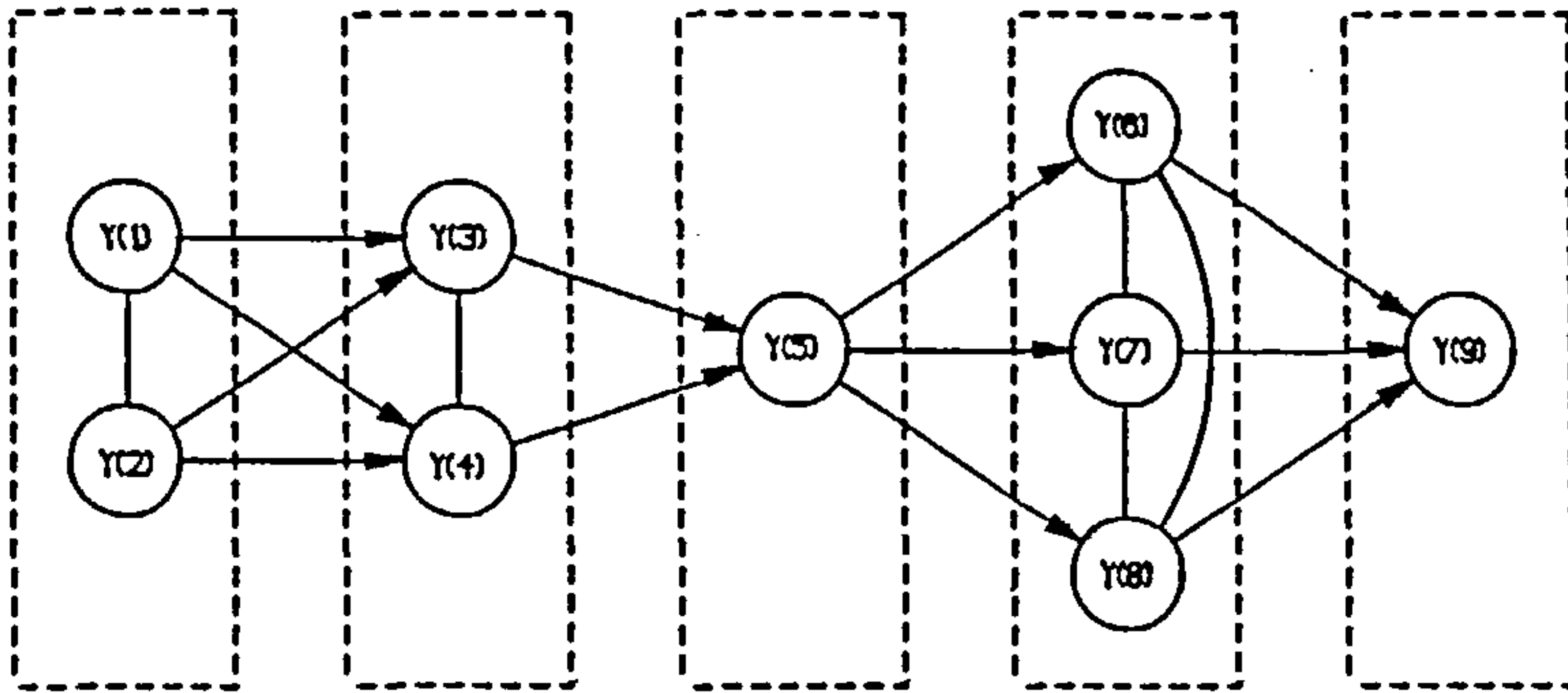


Figure 6.2: A research hypothesis of a real market with blocks $\{Y(1), Y(2)\}$, $\{Y(3), Y(4)\}$, $\{Y(5)\}$, $\{Y(6), Y(7), Y(8)\}$, $\{Y(9)\}$.

The DGM will be formally defined in section 6.2 and it will be shown how the model regressors follow a (conditionally normal) MDM. Section 6.3 introduces linear DGM's and shows how the joint forecast mean and covariance for these models is easily derived. Finally, in section 6.4 the models are demonstrated by a working example of the simplified shampoo market.

6.2 Dynamic Graphical Models.

Suppose that at a fixed time t , $N = \sum_i q_i$ variables are represented by a chain graph of the type described in the introduction and that the relationship between the N variables at time t and the past series is represented by a graph in which each variable has its previous series as parents and also the previous series of its parents at time t . List the components of the vector time series as $Y_t^T = (Y_t(1)^T, \dots, Y_t(n)^T)$ where each vector $Y_t(r)^T = (Y_{t1}(r), \dots, Y_{tq_r}(r))$ consists of the q_r variables in the r^{th} block of the chain. Conditional on the values of their parents, $Y_t(r)$ will follow a DLM if $q_r = 1$ and the symmetric DMR evolution if

$q_r \geq 2$. Suppose that $\Theta_t(r) = (\theta_{t1}(r), \dots, \theta_{tq_r}(r))$ is the state parameter matrix defining the distribution of $Y_t(r)$ and $\theta_{tj}(r)$ is the s_r dimensional state vector defining the distribution of $Y_{tj}(r)$, $1 \leq r \leq n$. Let Σ denote the $N \times N$ constant but unknown covariance matrix for Y_t whose r^{th} diagonal entry, $\Sigma(r)$, is the $q_r \times q_r$ covariance matrix for $Y_t(r)$.

As with the MDM, the DMR models may have *functions* of unobserved contemporary variables as regressors. Because of the symmetry of the roles played by the variables within a partition block, it will be assumed here that $Y_t(r)$ only has functions of *sums* of components over a parent block as regressors. Thus the sums are sufficient statistics for the blocks. Notice how this differs from MDM's where all the individual components of parent blocks would be regressors. For notational simplicity these sums will be labelled by:

$$\begin{aligned} X_t^*(r)^T &= \left(\sum_{j=1}^{q_1} Y_{tj}(1), \dots, \sum_{j=1}^{q_{r-1}} Y_{tj}(r-1) \right) \\ Z_t^*(r)^T &= \left(\sum_{j=1}^{q_{r+1}} Y_{tj}(r+1), \dots, \sum_{j=1}^{q_n} Y_{tj}(n) \right) \end{aligned}$$

such that

$$\begin{aligned} X^{*t}(r)^T &= \left(\sum_{j=1}^{q_1} Y_j^t(1), \dots, \sum_{j=1}^{q_{r-1}} Y_j^t(r-1) \right) \\ Z^{*t}(r)^T &= \left(\sum_{j=1}^{q_{r+1}} Y_j^t(r+1), \dots, \sum_{j=1}^{q_n} Y_j^t(n) \right) \end{aligned}$$

where $\sum_{j=1}^{q_r} Y_j^t(r)^T = (\sum_{j=1}^{q_r} Y_{1j}(r)^T, \dots, \sum_{j=1}^{q_r} Y_{tj}(r)^T)$. Both $X_t(r)$ and $Z_t(r)$ remain as defined by equation 5.1.

Explicitly, then, the DGM is given by:

Observation equations

$$Y_t(r)^T = F_t(r)^T \Theta_t(r) + v_t(r)^T, \quad 1 \leq r \leq n, \quad v_t(r) \sim N(\mathbf{0}, V_t(r)\Sigma(r)),$$

System equation

$$\Theta_t = G_t \Theta_{t-1} + \Omega_t, \quad \Omega_t \sim N(0, W_t, \Sigma),$$

Prior information

$$(\Theta_{t-1}, \Sigma | \mathbf{y}^{t-1}) \sim NW_{n_{t-1}}^{-1}(M_{t-1}, C_{t-1}, S_{t-1}')$$

$F_t(r)$ is an s_r dimensional vector and is allowed to be an arbitrary but known function of $\mathbf{x}^{*t}(r)$ and $\mathbf{y}^{t-1}(r)$, but *not* $\mathbf{y}^t \setminus \{\mathbf{y}^t(r)\}$, $\mathbf{z}^{*t}(r)$, $\sum_{j=1}^{q_r} \mathbf{y}_j^t(r)$ or $\mathbf{y}_t(r)$; $\mathbf{v}_t(r)$ is the q_r dimensional observation error vector; $V_t(r)$ is some known scalar;

$$\Theta_t = \text{blockdiag}(\Theta_t(1), \dots, \Theta_t(n))$$

$$\Omega_t = \text{blockdiag}(\Omega_t(1), \dots, \Omega_t(n))$$

and

$$M_{t-1} = \text{blockdiag}(M_{t-1}(1), \dots, M_{t-1}(n))$$

are all $(s \times N)$ matrices with $(s_r \times q_r)$ matrices on their diagonal such that Θ_t and Ω_t are the parameter matrix and system error matrix respectively and $\Omega_t(r)$ is the system error matrix for $Y_t(r)$; S_{t-1}' is a $(N \times N)$ matrix with the $(q_r \times q_r)$ matrices $\{S_{t-1}(1)n_{t-1}(1), \dots, S_{t-1}(n)n_{t-1}(n)\}$ on its diagonal where $n_{t-1}(1), \dots, n_{t-1}(n)$ are specified scalars; and

$$G_t = \text{blockdiag}(G_t(1), \dots, G_t(n)),$$

$$W_t = \text{blockdiag}(W_t(1), \dots, W_t(n))$$

and

$$C_{t-1} = \text{blockdiag}(C_{t-1}(1), \dots, C_{t-1}(n))$$

are all $(s \times s)$ matrices where $G_t(r)$, $W_t(r)$ and $C_{t-1}(r)$ are $(s_r \times s_r)$ square matrices which may be functions of past vectors $\mathbf{x}^{*t-1}(r)$ and $\mathbf{y}^{t-1}(r)$, but nothing else.

As for both the MDM and DMR models it is assumed that $\{v_{ij}(r)\}$ and $\{w_{ij}(r)\}$ are both independent for $1 \leq j \leq q_r$, $1 \leq r \leq n$, and are mutually independent over time $t \geq 1$.

By substituting $\Theta_t(r)$, $\{\mathbf{x}^{*t}(r), \mathbf{x}^t(r)\}$, $\{\sum_{j=1}^{q_r} \mathbf{y}_j^t(r), \mathbf{y}^t(r)\}$ and $\{\mathbf{z}^{*t}(r), \mathbf{z}^t(r)\}$ for $\theta_t(r)$, $\mathbf{x}^t(r)$, $\mathbf{y}^t(r)$ and $\mathbf{z}^t(r)$ respectively in Corollary 5.4.2 of section 5.4, the analogous form of equations 5.7 and 5.8 for DGM's follow directly. That is, if $\Pi_{r=1}^n \Theta_0(r)$, then for all time t :

$$\Pi_{r=1}^n \Theta_t(r) | \mathbf{y}^t, \sum_{j=1}^{q_1} \mathbf{y}_j^t(1), \dots, \sum_{j=1}^{q_n} \mathbf{y}_j^t(n) \quad (6.1)$$

and

$$\Theta_t(r) \Pi \mathbf{z}^t(r), \mathbf{z}^{*t}(r) | \mathbf{x}^t(r), \mathbf{x}^{*t}(r), \mathbf{y}^t(r), \sum_{j=1}^{q_r} \mathbf{y}_j^t(r) \quad (6.2)$$

So, as with MDM's, the conditional distribution of $Y_t(r) | \{Y_t(1), \dots, Y_t(r-1)\}$, $1 \leq r \leq n$ for each block can be forecast separately and updated in closed form (by equation 6.1). Once again, the joint forecast distribution can then be expressed as the product of these conditional distributions and is given by:

$$p\{\mathbf{y}_t | \mathbf{y}^{t-1}\} = \prod_r \int_{\Theta_t(r)} p\{\mathbf{y}_t(r) | \mathbf{x}_t^*(r), \mathbf{y}^{t-1}(r), \Theta_t(r)\} p\{\Theta_t(r) | \mathbf{y}^{t-1}\} d\Theta_t(r).$$

Now, in the definition of the DGM for Y_t , it states that $G_t(r)$, $W_t(r)$ and $C_{t-1}(r)$ may be functions of $\mathbf{y}^{t-1}(r)$ and $\mathbf{x}^{*t-1}(r)$ but nothing else. This, together with equation 6.2, allows the simplification of $p\{\Theta_t(r) | \mathbf{y}^{t-1}\}$ to:

$$p\{\Theta_t(r) | \mathbf{y}^{t-1}\} = p\{\Theta_t(r) | \mathbf{x}^{*t-1}(r), \mathbf{y}^{t-1}(r)\}.$$

Therefore it is only necessary that the series of *sums* of the components in each of the first $r - 1$ blocks are known to find the conditional forecast distribution for $Y_t(r)$.

It is clear that, given the additional form of $F_t(r)$, the conditional causal relationships found in a DGM differ from those of the MDM in that in the DGM $Z(r)$, $Z^*(r)$, $X(r)$ and $\sum_{j=1}^{q_r} y_j(r)$ are all non-causes of $Y(r)$ given $X^*(r)$.

It is interesting to note that $X_t^{*T} = (\sum_{j=1}^{q_1} Y_t(1), \dots, \sum_{j=1}^{q_n} Y_t(n))$ are governed by a (conditionally normal) MDM whose conditional variances are estimated on-line through the estimation of Σ . Suppose that 1_{q_r} is a q_r dimensional vector such that $1_{q_r}^T = (1, \dots, 1)$. The MDM across these regressors is then given by:

Observation equation

$$\sum_{j=1}^{q_r} Y_{tj}(r) = F_t(r)^T \Theta_t(r) 1_{q_r} + v_t(r)^T 1_{q_r}, \quad v_t(r)^T 1_{q_r} \sim N(0, V_t(r) \Sigma^*(r))$$

$$1 \leq r \leq n$$

System equation

$$\Theta_t(r) 1_{q_r} = G_t(r) \Theta_{t-1}(r) 1_{q_r} + \Omega_t(r) 1_{q_r}, \quad \Omega_t(r) 1_{q_r} \sim N(\mathbf{o}, \Sigma^*(r) W_t(r))$$

$$1 \leq r \leq n$$

Prior information

$$(\Theta_{t-1}(r) 1_{q_r} | y^{t-1}, \Sigma^*(r)) \sim N(M_{t-1}(r) 1_{q_r}, C_{t-1}(r) \Sigma^*(r))$$

$$(\Sigma^*(r)^{-1} | y^{t-1}) \sim G\left(\frac{n_{t-1}(r)}{2}, \frac{S_{t-1}^*(r) n_{t-1}(r)}{2}\right)$$

where $\Sigma^*(r) = 1_{q_r}^T \Sigma(r) 1_{q_r}$, $S_{t-1}^*(r) = 1_{q_r}^T S_{t-1}(r) 1_{q_r}$ and $F_t(r)$ and $G_t(r)$ are the same as in the DGM.

Proof: From multivariate normal theory (see for example Chatfield & Collins, 1980), it is known that if

$$v_t(r)^T \sim N(\mathbf{o}^T, V_t(r) \Sigma(r))$$

then

$$v_t(r)^T 1_{q_r} \sim N(\mathbf{o}^T 1_{q_r}, V_t(r) 1_{q_r}^T \Sigma(r) 1_{q_r})$$

and so it has been proved that $v_t(r)^T 1_{q_r}$ has a univariate normal distribution given by:

$$v_t(r)^T 1_{q_r} \sim N(0, V_t(r) \Sigma^*(r)).$$

It will now be proved that:

$$\Omega_t(r) 1_{q_r} \sim N(\mathbf{o}, \Sigma^*(r) W_t(r)).$$

From matrix-variate normal theory (see for example Dawid, 1981) there is a result which states that if:

$$V \sim N(A, B, C)$$

then

$$HVK + L \sim N(HAK + L, HBH^T, K^TCK). \quad (6.3)$$

Therefore, since

$$\Omega_t(r) \sim N(0, W_t(r), \Sigma(r))$$

then

$$\Omega_t(r) 1_{q_r} \sim N(0 1_{q_r}, W_t(r), 1_{q_r} \Sigma(r) 1_{q_r}) \equiv N(\mathbf{o}, W_t(r), \Sigma^*(r))$$

where \mathbf{o} is now an s_r -dimensional vector. The matrix normal density of $\Omega_t(r) 1_{q_r}$ is therefore given by:

$$p\{\Omega_t(r) 1_{q_r}\} = k \exp \left\{ -\frac{1}{2} \text{trace} \left[\{\Omega_t(r) 1_{q_r}\}^T W_t(r)^{-1} \{\Omega_t(r) 1_{q_r}\} \Sigma^*(r)^{-1} \right] \right\}$$

where

$$k = (2\pi)^{-s_r/2} |W_t(r)|^{-1/2} |\Sigma^*(r)|^{-s_r/2}.$$

Since $\Sigma^*(r)$ is a scalar and the trace of a scalar is merely that scalar, $p\{\Omega_t(r)1_{q_r}\}$ becomes:

$$p\{\Omega_t(r)1_{q_r}\} = k \exp \left\{ -\frac{1}{2} \left[\{\Omega_t(r)1_{q_r}\}^T \{\Sigma^*(r)W_t(r)\}^{-1} \{\Omega_t(r)1_{q_r}\} \right] \right\}.$$

Now, the determinant of a scalar is also simply that scalar and so the constant term becomes:

$$\begin{aligned} k &= (2\pi)^{-s_r/2} |W_t(r)|^{-1/2} \Sigma^*(r)^{-s_r/2} \\ &= (2\pi)^{-s_r/2} \left\{ \Sigma^*(r)^{-\frac{1}{2}} \right\}^{s_r} |W_t(r)|^{-\frac{1}{2}} \end{aligned}$$

which by the properties of determinants becomes:

$$k = (2\pi)^{-\frac{s_r}{2}} |\Sigma^*(r)W_t(r)|^{-\frac{1}{2}}.$$

Therefore

$$\begin{aligned} p\{\Omega_t(r)1_{q_r}\} &= (2\pi)^{-\frac{s_r}{2}} |\Sigma^*(r)W_t(r)|^{-\frac{1}{2}} \\ &\times \exp \left\{ -\frac{1}{2} \left[\{\Omega_t(r)1_{q_r}\}^T \{\Sigma^*(r)W_t(r)\}^{-1} \{\Omega_t(r)1_{q_r}\} \right] \right\} \end{aligned}$$

which is a multivariate normal density, thus proving that:

$$\Omega_t(r)1_{q_r} \sim N(\mathbf{o}, \Sigma^*(r)W_t(r)).$$

The distributions which make up the prior information will now be derived. Throughout the derivation assume that all distributions are conditional on D_0 . It is known that:

$$(\Theta_0(r) | \Sigma(r)) \sim N(M_0(r), C_0(r), \Sigma(r)).$$

By equation 6.3 the result comes directly that:

$$(\Theta_0(r)1_{q_r} | \Sigma(r)) \sim N(M_0(r)1_{q_r}, C_0(r), \Sigma^*(r))$$

and following the same reasoning as in the last proof, the derivation of the distribution of $\{\Theta_0(r)1_{q_r} | \Sigma(r)\}$ can be completed to give:

$$(\Theta_0(r)1_{q_r} | \Sigma(r)) \sim N(M_0(r)1_{q_r}, \Sigma^*(r)C_0(r)).$$

It now remains to derive the initial distribution of $\Sigma^*(r)$. From the DGM it is known that:

$$\Sigma(r)^{-1} \sim W_{n_0(r)}(S_0(r)n_0(r)).$$

From Wishart distribution theory (see for example Chatfield & Collins, 1990) this implies that:

$$\{1_{q_r}^T \Sigma(r) 1_{q_r}\}^{-1} \{n_0(r) 1_{q_r}^T S_0(r) 1_{q_r}\} \sim \chi_{n_0(r)}^2.$$

By the relationship between the χ^2 and gamma distributions, this becomes:

$$\Sigma^*(r)^{-1} \sim G\left(\frac{n_0(r)}{2}, \frac{S_0^*(r)n_0(r)}{2}\right)$$

so that $S_0^*(r)$ is the point estimate of $\Sigma^*(r)$ at time t .

6.3 Linear Dynamic Graphical Models.

Section 5.5 introduced linear MDM's and corrected linear MDM's which are particularly simple to work with. DGM's have two analogous models — namely the *linear DGM's* (LDGM's) and *corrected linear DGM's* (CLDGM's). They are a combination of the LMDM's/CLMDM's and the DMR model and as such inherit the simplicity of these models.

Suppose that $\{v_t(r), \Omega_t(r)\}_{t \geq 1}$ are jointly Gaussian, are independent of y^{t-1} and $F_t(1)$ does not depend on Y_t . The LDGM is defined so that

$$F_t(r)^T = (x_t^*(r)^T, \tilde{x}_t(r)^T)$$

where $\tilde{\mathbf{x}}_t(r)$ is a set of known exogenous variables, and the CLDGM sets

$$F_t(r)^T = \left[\left\{ \mathbf{x}_t^*(r) - \hat{\mathbf{f}}_t^*(r) \right\}^T, \tilde{\mathbf{x}}_t(r)^T \right]$$

where $\hat{\mathbf{f}}_t^*(r)^T = \left(E \left[\sum_{j=1}^{q_1} y_{tj}(1) \mid \mathbf{y}^{t-1} \right], \dots, E \left[\sum_{j=1}^{q_{r-1}} y_{tj}(r-1) \mid \mathbf{y}^{t-1} \right] \right)$.

As for the LMDM and CLMDM, the individual forecast distributions of the components $Y_t(r)$, $r = 1, \dots, n$, are multivariate normal or T. The joint forecast distribution is then simply the product of the conditional distributions $\{Y_t(r) \mid \mathbf{x}_t^*(r)\}$, $r = 1, \dots, n$. Once again, the joint forecast distribution of Y_t is not of a simple form but, as for LMDM's and CLMDM's the means and variances/covariances of the joint forecast distribution can be found explicitly fairly easily and these will be derived now.

The conditional forecast mean of $\{Y_t(r) \mid \mathbf{x}_t^*(r)\}$ when each $\mathbf{y}_t(r)$ is modelled by an LDGM comes directly from equation 3.13 so that:

$$E\{Y_t(r)^T \mid \mathbf{y}^{t-1}(r), \mathbf{x}^{*t}(r)\} = F_t(r)^T a_t(r)$$

where $a_t(r)$ is an $(s_r \times q_r)$ matrix. Let $a_t(r)$ be partitioned in a similar way to the analogous LMDM and CLMDM, so that

$$a_t(r)^T = (A_t^*(r)^T, \tilde{A}_t(r)^T)$$

where

$$A_t^*(r) = (a_{t1}(r), \dots, a_{tq_r}(r))$$

such that the $(r-1)$ -dimensional vector $a_{tj}(r)^T = (a_{tj}^{(1)}(r), \dots, a_{tj}^{(r-1)}(r))$ contains those parameters associated with $\mathbf{x}_t^*(r)$ and $\tilde{A}_t(r)$ is the $((s_r - r + 1) \times q_r)$ matrix of parameters associated with $\{\tilde{\mathbf{x}}_t(r), \mathbf{y}^{t-1}(r), \mathbf{x}^{*t-1}(r)\}$. The conditional forecast expectation of $\{Y_t(r) \mid \mathbf{x}_t^*(r)\}$ can then be expressed by:

$$\left. \begin{aligned} E\{Y_t(1) \mid \mathbf{y}^{t-1}(1)\} &= a_t^{(0)}(r) \\ E\{Y_t(r) \mid \mathbf{y}^{t-1}(r), \mathbf{x}^{*t}(r)\} &= a_t^{(0)}(r) + A_t(r)^T \mathbf{x}_t^*(r) \end{aligned} \right\}$$

where $\mathbf{a}_i^{(0)}(r)$ is a q_r dimensional vector which is a function of $\tilde{A}_i(r)$, $\tilde{\mathbf{x}}_i(r)$, $\mathbf{y}^{t-1}(r)$, $\mathbf{x}^{*t-1}(r)$ only.

By equation 5.21 the mean of the joint forecast distribution of \mathbf{Y}_t can be found directly to give:

$$E[\mathbf{Y}_t | \mathbf{y}^{t-1}] = \mathbf{a}_i^{(0)} + A_t E[\mathbf{X}_i^* | \mathbf{y}^{t-1}]$$

where $\mathbf{a}_i^{(0)}$ is an N dimensional vector such that:

$$\mathbf{a}_i^{(0)T} = \left(\mathbf{a}_i^{(0)}(1)^T, \dots, \mathbf{a}_i^{(0)}(n)^T \right)$$

and the lower triangular $N \times n$ matrix A_t is such that the $(q_i + 1)^{th}$ row to the q_{i+1}^{th} row has $A_t(i)$ in its first $i - 1$ columns.

Now, since \mathbf{X}_i^* will follow an LMDM, then from section 5.5 it is known that:

$$E \left\{ \sum_{j=1}^{q_r} Y_{tj}(r) | \mathbf{y}^{t-1}, \mathbf{x}^{*t}(r) \right\} = \mathbf{a}_i^{(0)\Sigma}(r) + \left[\sum_{i=1}^{r-1} \left\{ \mathbf{a}_i^{(i)\Sigma}(r) \sum_{j=1}^{q_i} Y_{tj}(i) \right\} \right]$$

where $\mathbf{a}_i^{(0)\Sigma}(r) = \sum_{j=1}^{q_r} \mathbf{a}_{tj}^{(0)}(r)$ and $\mathbf{a}_i^{(i)\Sigma}(r) = \sum_{j=1}^{q_r} \mathbf{a}_{tj}^{(i)}(r)$. Thus,

$$E\{\mathbf{X}_i^* | \mathbf{y}^{t-1}\} = [I - A^\Sigma]^{-1} \mathbf{a}_i^{(0)\Sigma}$$

where A^Σ is an $n \times n$ lower triangular matrix whose $(j, k)^{th}$ element a_{jk}^Σ is given by:

$$a_{jk}^\Sigma = \begin{cases} \mathbf{a}_i^{(k)\Sigma}(j) & k < j \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbf{a}_i^{(0)\Sigma} = \left(\mathbf{a}_i^{(0)\Sigma}(1), \dots, \mathbf{a}_i^{(0)\Sigma}(n) \right)^T.$$

Thus the expectation of the joint forecast distribution of $\{\mathbf{Y}_t | \mathbf{y}^{t-1}\}$ has been found explicitly.

Similarly, when \mathbf{Y}_t is modelled by a CLDGM the expectation of the conditional forecast distribution of $\{\mathbf{Y}_t(r) | \mathbf{x}_i^*(r)\}$ is given by:

$$E[\mathbf{Y}_t(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^{*t}(r)] = \bar{\mathbf{a}}_i^{(0)}(r) + \bar{A}_t(r)^T \{\mathbf{x}_i^*(r) - \hat{\mathbf{f}}_i^*(r)\} \quad (6.4)$$

where the q_r dimensional vector $\bar{\mathbf{a}}_i^{(0)}(r)$ and the $(r-1) \times q_r$ matrix $\bar{\mathbf{A}}_i(r) = (\bar{\mathbf{a}}_{i1}(r), \dots, \bar{\mathbf{a}}_{iq_r}(r))$ are similar to $\mathbf{a}_i^{(0)}(r)$ and $\mathbf{A}_i(r)$ but with different values. The marginal mean for Y_i is similar to the analogous result for CLMDM's and is simply:

$$E[Y_i | \mathbf{y}^{t-1}] = \bar{\mathbf{a}}_i^{(0)}. \quad (6.5)$$

Finding the forecast covariance for Y_i is a little more complicated. The workings for both the linear and corrected linear models follow the same principles, but as is the case for MDM's, those for the corrected linear model are more straightforward and so only the derivation of the forecast covariance of that model will be presented here. All workings are shown unconditionally on knowing Σ and so estimates of this matrix are substituted throughout.

By equation 3.13 the conditional forecast covariance of $\{Y_i(r) | \mathbf{x}_i^*(r)\}$, $1 \leq r \leq n$, is given by:

$$\text{cov}\{Y_i(r) | \mathbf{y}^{t-1}(r), \mathbf{x}^{*t}(r)\} = S_{t-1}(r) \{V_i(r) + F_i(r)^T R_i(r) F_i(r)\}.$$

Suppose that each of the N $\theta_{ij}(r)$, $1 \leq j \leq q_r$, $1 \leq r \leq n$, can be written as:

$$\theta_{ij}(r) = (\theta_{ij}^*(r), \tilde{\theta}_{ij}(r))$$

where $\theta_{ij}^*(r)$ is the set of parameters in $\theta_{ij}(r)$ associated with $\mathbf{x}_i^*(r)$ and $\tilde{\theta}_{ij}(r)$ is the set associated with $\tilde{\mathbf{x}}_i(r)$. If $R_i(r)$ is the covariance matrix for $\{\theta_{ij}(r) | \mathbf{y}^{t-1}\}$, for each j , $1 \leq j \leq q_r$, then $R_i(r)$ can be expressed by:

$$R_i(r) = \begin{pmatrix} R_i^*(r) & R_i'(r) \\ R_i'(r)^T & \tilde{R}_i(r) \end{pmatrix}$$

where for each $j = 1, \dots, q_r$

$$R_i^*(r) = \text{cov}(\theta_{ij}^*(r), \theta_{ij}^*(r))$$

$$R_i'(r) = \text{cov}(\theta_{ij}^*(r), \tilde{\theta}_{ij}(r))$$

$$\tilde{R}_i(r) = \text{cov}(\tilde{\theta}_{ij}(r), \tilde{\theta}_{ij}(r))$$

The conditional forecast covariance can then be rewritten as:

$$\left. \begin{aligned} \text{cov}\{Y_t(1) | y^{t-1}(1)\} &= \tau_t^2(1) \\ \text{cov}\{Y_t(r) | y^{t-1}(r), x^{*t}(r)\} &= S_{t-1}(r)D_t(r), \end{aligned} \right\} \quad 2 \leq r \leq n$$

such that:

$$D_t(r) = \left[\tau_t^2(r) + \{x_t^*(r) - \hat{f}_t^*(r)\}^T R_t^*(r) \{x_t^*(r) - \hat{f}_t^*(r)\} \right]$$

where $\tau_t^2(r)$ is a function of $\{V_t(r), \tilde{x}_t(r), R_t'(r), \tilde{R}_t(r)\}$. Following the same argument presented in section 5.5, this can be rewritten as:

$$\text{cov}(Y_t(r) | y^{t-1}(r), x^{*t}(r)) = S_{t-1}(r)Q_t(r) \quad (6.6)$$

where

$$Q_t(r) = \left[\tau_t^2(r) + \text{trace} \left[U_t(r)^T \{x_t^*(r) - \hat{f}_t^*(r)\} \{x_t^*(r) - \hat{f}_t^*(r)\}^T U_t(r) \right] \right]$$

such that $R_t^*(r) = U_t(r)U_t(r)^T$ and $U_t(r)$ is a singular matrix.

It is relatively straight forward to derive the covariance matrix of the forecast distribution of $Y_t(r)$, unconditionally on $x_t^*(r)$. Let $\bar{\Sigma}_t(r)$ be the covariance matrix for the forecast distribution of $\{X_t^*(r) | y^{t-1}\}$. From the identity 5.25 of section 5.5, together with equations 6.4 and 6.6 it is clear that the covariance of the forecast distribution of $Y_t(r)$ is given by:

$$\text{cov}\{Y_t(r) | y^{t-1}(r), x^{*t-1}(r)\} = S_{t-1}(r)E[Q_t(r)] + \bar{A}_t(r)^T \bar{\Sigma}_t(r) \bar{A}_t(r)$$

where

$$E[Q_t(r)] = \tau_t^2(r) + \text{trace}\{U_t(r)^T \bar{\Sigma}_t(r) U_t(r)\}.$$

Now the covariance matrix between $Y_t(k)$ and $Y_t(l)$, $k \neq l$, $k, l = 1, \dots, n$ will be found. As was mentioned in section 6.2, $Z(r)$, $Z^*(r)$, $X(r)$ and $\sum_{j=1}^{q_r} Y_j(r)$ are all non-causes of $Y(r)$ so that:

$$Y_t(r) \perp\!\!\!\perp x^{t-1}(r), z^{t-1}(r), z^{*t-1}(r), \sum_{j=1}^{q_r} y_j^{t-1}(r) | y^{t-1}(r), x^{*t}(r).$$

Using this conditional causal relationship, together with the form of $G_t(r)$ and $F_t(r)$ it can be concluded that for $k \neq l$, $k, l = 1, \dots, n$:

$$\text{cov} \{Y_t(k), Y_t(l) | \mathbf{x}^{*t}(k), \mathbf{x}^{*t}(l), \mathbf{y}^{t-1}\} = 0.$$

Therefore

$$E \{Y_t(k)Y_t(l) | \mathbf{x}^{*t}, \mathbf{y}^{t-1}\} = E \{Y_t(k) | \mathbf{x}^{*t}, \mathbf{y}^{t-1}\} E \{Y_t(l) | \mathbf{x}^{*t}, \mathbf{y}^{t-1}\}.$$

Let $\bar{\Sigma}_t$ be the forecast covariance matrix of X_t^* such that:

$$\{\bar{\Sigma}_t\}_{kl} = \bar{\sigma}_t(k, l) = \text{cov} \left\{ \sum_{j=1}^{q_k} Y_{tj}(k), \sum_{j=1}^{q_l} Y_{tj}(l) | \mathbf{y}^{t-1} \right\}, \quad 1 \leq j, k \leq n$$

and let $\bar{\Sigma}_t(k, l)$ be the $(k-1) \times (l-1)$ forecast covariance matrix of $X_t^*(k)$ and $X_t^*(l)$. As X_t^* follows a LMDM $\bar{\Sigma}_t$ and $\bar{\Sigma}_t(k, l)$ can be found explicitly by using the methodology of section 5.5. Using the identity 5.21 of section 5.5 and equation 6.4 after some algebra it is clear that:

$$E \{Y_t(k)Y_t(l) | \mathbf{y}^{t-1}, \mathbf{x}^{*t-1}\} = \bar{\mathbf{a}}_t^{(0)}(k)\bar{\mathbf{a}}_t^{(0)}(l)^T + \bar{\mathbf{A}}_t(k)^T \bar{\Sigma}_t(k, l) \bar{\mathbf{A}}_t(l). \quad (6.7)$$

It is already known that:

$$E \{Y_t(i) | \mathbf{y}^{t-1}, \mathbf{x}^{*t-1}\} = \bar{\mathbf{a}}_t^{(0)}(i), \quad 1 \leq i \leq n$$

from equation 6.5, so that:

$$E \{Y_t(k) | \mathbf{y}^{t-1}, \mathbf{x}^{*t-1}\} E \{Y_t(l) | \mathbf{y}^{t-1}, \mathbf{x}^{*t-1}\} = \bar{\mathbf{a}}_t^{(0)}(k)\bar{\mathbf{a}}_t^{(0)}(l).$$

By subtracting this from equation 6.7 it follows immediately that the covariance of the joint forecast distribution of Y_t , when Y_t follows a CLDGM is given by:

$$\text{cov} \{Y_t(k), Y_t(l) | \mathbf{y}^{t-1}, \mathbf{x}^{*t-1}\} = \bar{\mathbf{A}}_t(k)^T \bar{\Sigma}_t(k, l) \bar{\mathbf{A}}_t(l).$$

Thus once $\bar{\Sigma}_t$ and $\bar{\Sigma}_t(k, l)$ have been found by using the recursive relationships defined in section 5.5, the joint forecast covariance matrix of Y_t can be found explicitly.

6.4 A Simple Illustration of a DGM.

To illustrate the consequences of using a particular DGM, return to the original shampoo example. Let $Y_i(1)$ and $Y_i(3)$ be the blocks I_i and S_i respectively and $Y_i(2)^T = (Y_{i1}(2), Y_{i2}(2))$ represent the block $\{C_i, N_i\}$. Suppose that $\Sigma(1)$, $\Sigma(2)$ and $\Sigma(3)$ are the unknown variances/covariances of $Y(1)$, $Y(2)$ and $Y(3)$ respectively. The linear DGM for this example is given by the following observation and system equations. Notice that $y_i(1)$ is a regressor in $Y_i(2)$'s model, since $Y_i(1)$ is a parent of $Y_i(2)$; and the sum $y_{i1}(2) + y_{i2}(2)$ is a regressor in $Y_i(3)$'s model, since $Y_i(2)$ is a parent block of $Y_i(3)$.

Observation equations

$$\begin{aligned} Y_i(1) &= \theta_i^{(0)}(1) + v_i(1), & v_i(1) &\sim N(0, \Sigma(1)) \\ Y_i(2)^T &= (1, y_i(1)) \Theta_i(2) + v_i(2)^T, & v_i(2) &\sim N(\mathbf{o}, \Sigma(2)) \\ Y_i(3) &= (1, y_{i1}(2) + y_{i2}(2)) \theta_i(3) + v_i(3), & v_i(3) &\sim N(0, \Sigma(3)) \end{aligned}$$

System equations

$$\begin{aligned} \theta_i^{(0)}(1) &= \theta_{i-1}^{(0)}(1) + w(1)_i, & w(1)_i &\sim N(0, W_i(1)\Sigma(1)) \\ \Theta_i(2) &= \Theta_{i-1}(2) + \Omega_i(2), & \Omega_i(2) &\sim N(0, W(2), \Sigma(2)) \\ \theta_i(3) &= \theta_{i-1}(3) + w_i(3), & w_i(3) &\sim N(\mathbf{o}, W(3)\Sigma(3)) \end{aligned}$$

Prior Information

$$\begin{aligned} (\theta_i^{(0)}(1) | D_0, \Sigma(1)) &\sim N(m_0(1), C_0(1)\Sigma(1)) \\ (\Sigma(1)^{-1} | D_0) &\sim G\left(\frac{n_0(1)}{2}, \frac{S_0(1)n_0(1)}{2}\right), \\ (\Theta_0(2), \Sigma(2) | D_0) &\sim NW_{n_0(2)}^{-1}(M_0(2), C_0(2), S_0(2)n_0(2)), \\ (\theta_0(3) | D_0, \Sigma(3)) &\sim N(m_0(3), C_0(3)\Sigma(3)) \end{aligned}$$

$$(\Sigma(3)^{-1} | D_0) \sim G\left(\frac{n_0(3)}{2}, \frac{S_0(3)n_0(3)}{2}\right),$$

where

$$\Theta_t(2) = \begin{pmatrix} \theta_{t1}^{(0)}(2) & \theta_{t2}^{(0)}(2) \\ \theta_{t1}^{(1)}(2) & \theta_{t2}^{(1)}(2) \end{pmatrix},$$

$\theta_t(3)^T = (\theta_t^{(0)}(3) \ \theta_t^{(1)}(3))$ and D_0 represents the knowledge of the system before any observations.

The monthly data set analysed is a gross modification of a non-seasonal market with 3 competitors when there was a sudden increase in the interest rate at period 13 (all other sources of variation have been filtered). Like MDM's, DGM's retain nearly all the advantages of univariate DLM's and, in particular, the familiar technique of intervention analysis (see section 3.4) can be used at this time period on series I_t . The conditional forecast distributions for each series are found easily from equations 3.8 and 3.13 and the one-step ahead conditional forecasts for each distribution are given in figure 6.3 (a). Notice that some movement in $Y(3)$ has occurred after period 13, but less than in $Y(2)$, reflecting the fact that secondary effects respond less strongly than primary effects. Notice from the non-elliptical contour plots in figure 6.4, that, as for the MDM, the joint forecast density is non-Gaussian, even in this very simple linear DGM. This process, therefore, is very different from a multivariate normal time series.

Using the notation derived in section 6.3 the marginal moments of components in blocks can be derived so that after a little algebra:

$$E\{Y_t(2)^T | \mathbf{y}^{t-1}(2)\} = [1, E\{Y_t(1) | \mathbf{y}^{t-1}(1)\}] M_{t-1}(2),$$

$$E\{Y_t(3) | \mathbf{y}^{t-1}(3)\} = [1, E\{Y_{t1}(2) + Y_{t2}(2) | \mathbf{y}^{t-1}(2)\}] m_{t-1}(3),$$

$$\text{var}\{Y_{tj}(2) | \mathbf{y}^{t-1}(2)\} = S_t(2)^{(j,j)} E\{Q_t(2)\} + [M_{t-1}(2)^{(2,j)}]^2 \text{var}[Y_t(1) | \mathbf{y}^{t-1}(1)],$$

$$j = 1, 2,$$

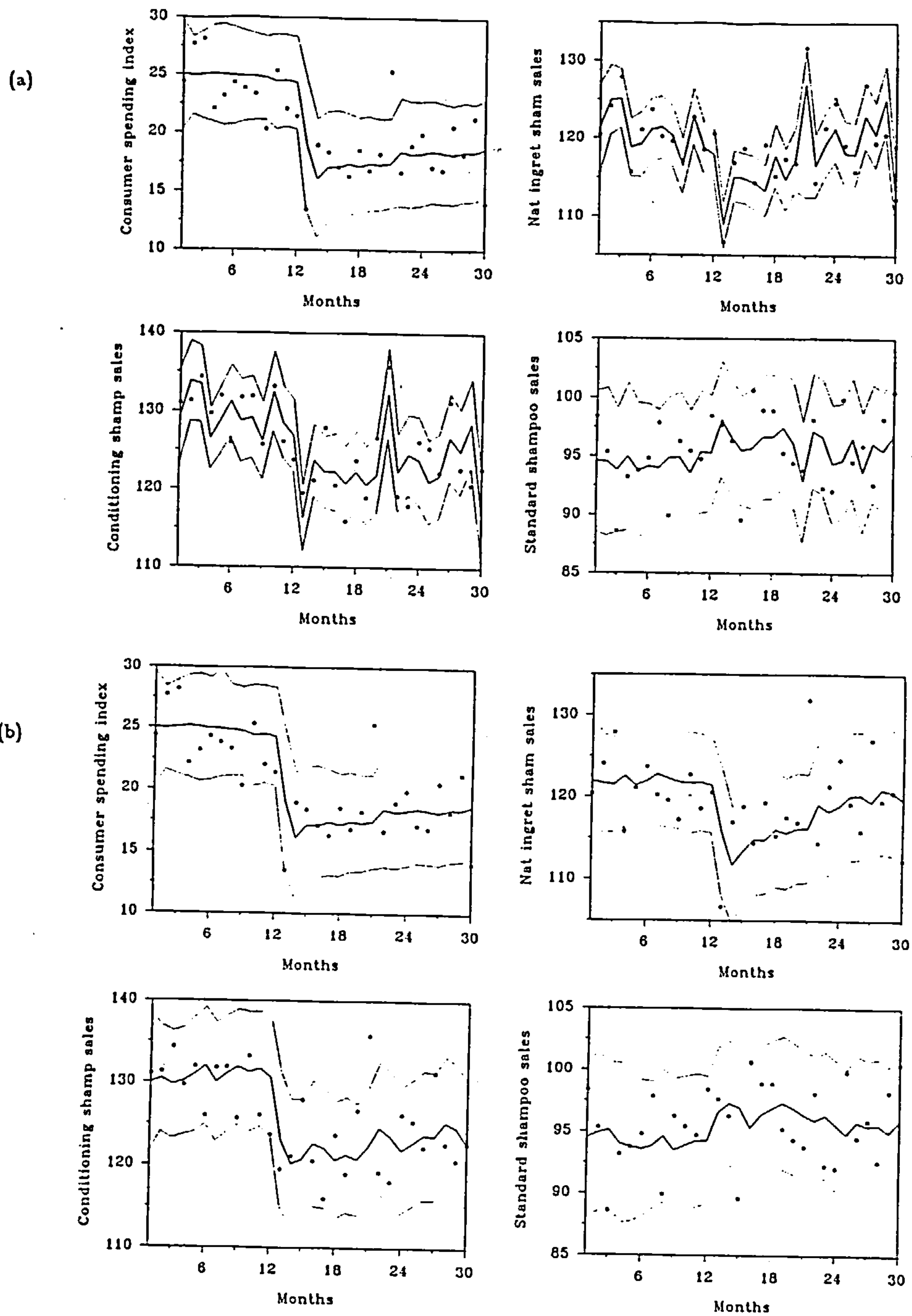


Figure 6.3: Conditional (a) and marginal (b) one-step-ahead forecasts of brand sales and index of consumer income in a shampoo market. The dots are the observations, the solid line gives the one-step ahead forecasts and the dotted lines represent the 90% confidence limits.

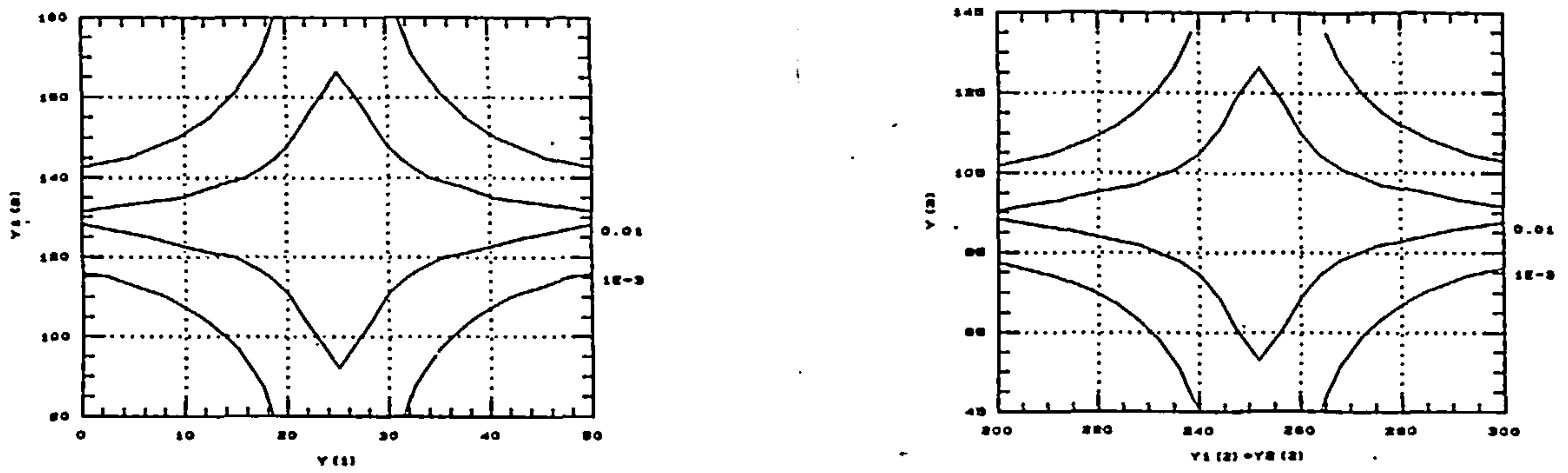


Figure 6.4: Contour plots of initial joint forecast densities of (a) $Y(1)$ and $Y_1(2)$, and (b) $Y_1(2) + Y_2(2)$ and $Y(3)$.

$$\begin{aligned} \text{var} \{Y_t(3) | \mathbf{y}^{t-1}(3)\} &= S_t(3)E\{Q_t(3)\} \\ &\quad + [m_{t-1}(3)^{(2)}]^2 \text{var} \{Y_{t1}(2) + Y_{t2}(2) | \mathbf{y}^{t-1}(2)\}, \end{aligned}$$

where:

$$\begin{aligned} E\{Q_t(2)\} &= 1 + R_t(2)^{(1,1)} + 2R_t(2)^{(1,2)}E[Y_t(1) | \mathbf{y}^{t-1}(1)] \\ &\quad + R_t(2)^{(2,2)} \left\{ \text{var} [Y_t(1) | \mathbf{y}^{t-1}(1)] + E[Y_t(1) | \mathbf{y}^{t-1}(1)]^2 \right\}, \\ E\{Q_t(3)\} &= 1 + R_t(3)^{(1,1)} + 2R_t(3)^{(1,2)}E[Y_{t1}(2) + Y_{t2}(2) | \mathbf{y}^{t-1}(2)] \\ &\quad + R_t(3)^{(2,2)}E[\{Y_{t1}(2) + Y_{t2}(2)\}^2 | \mathbf{y}^{t-1}(2)] \end{aligned}$$

and where the entry in the i^{th} row and j^{th} column of any matrix J is denoted by $J^{(i,j)}$ and the i^{th} entry of any vector K is denoted by $K^{(i)}$. The one-step ahead marginal forecasts can be seen in figure 6.3 (b).

6.5 Conclusion

Although the analysis of complex market structures with causal hypotheses using DGM's is in its infancy, preliminary results on real series are very encouraging. We believe that in the future they will extend the scope of the proven versatile DLM. In particular, they enable plausible strong and very unsymmetrical prior

information to be incorporated in multivariate processes. This permits strong inferences about the development of the process to be made without requiring very large uncontaminated series, virtually unavailable in the business environment.

Chapter 7

Partial Segmentation Models.

7.1 Introduction.

This chapter introduces a new class of Bayesian models which are a generalisation of the Dirichlet model and extend the work of Dickey et al (1987) to the study of models with latent conditional independence structures.

Recall the Dirichlet model of section 2.4 of chapter 2 for homogeneous markets. Here, the number of purchases of the various brands, represented by the vector τ , followed a multinomial distribution (see equation 2.2) and a Dirichlet prior was placed across the parameters ψ (see equation 2.1) where for each consumer in the market:

$$\psi_j = P(\text{purchase brand } j).$$

Suppose that in a certain partially segmented market there are m brands and the consumers are divided into n types where:

$$\theta(i) = P(\text{type } i).$$

In this case, individuals of different types do not necessarily have the same probability of purchasing brand j so that:

$$\psi_j = \sum_i P(\text{buy brand } j | \text{type } i) \theta(i)$$

and it was shown how the Dirichlet model is inappropriate for this type of situation. With only one m -dimensional observation vector \mathbf{r} the problem is clearly over-parameterised. A model is therefore required which focuses on the two sets of probabilities $P(\text{buy brand } j \mid \text{type } i)$ and $\theta(i)$, for $1 \leq i \leq n$, $1 \leq j \leq m$, but can allow consistent estimation of the probabilities ψ_j , $1 \leq j \leq m$.

Following Dickey et al. (1987), it is shown in section 7.2 how, given a hypothesised matrix Z of likelihood ratios and an observation vector \mathbf{r} , the likelihood on the purchase probabilities ψ separates. For consistency with the work of Dickey, “brands” will be known as *records* and “types” will be referred to as *outcomes* for the rest of this chapter. The likelihood of ψ then separates into a likelihood on the outcome probability parameter θ and the sample distribution scaling parameter λ , where

$$\lambda_j = \max_{1 \leq i \leq n} P(\text{record } j \mid \text{outcome } i).$$

Furthermore, the dimension of the (θ, λ) parameter vector is $m - 1$ and so consistent estimation of (θ, λ) , given the m -dimensional observation vector \mathbf{r} , is at least plausible.

In section 7.3 it is shown how a closed form prior to posterior analysis can be performed on the sample distribution scaling parameter λ with a certain large class of conditioning matrix Z . These lie in an apparently novel class of distributions which are called nested generalised Dirichlets. Within a certain subclass, these distributions can, in fact, be transformed by a reparameterisation into products on independent Dirichlets which are obviously particularly simple to analyse. These are discussed in section 7.4.

Models with local independence structures are often difficult to understand because their structure and the implications are hidden under a canopy of complex notation and side conditions. It is shown in section 7.5 how the graphical methods

of Lauritzen & Spiegelhalter (1988) can be used to represent these structures by undirected graphs which can then form the framework of a prior to posterior analysis. An intriguing relationship between the useful class of Z mentioned above and decomposable graphs is established. Section 7.6 shows how such graphs can be used to guide the parameterisation of λ into a convenient form.

7.2 Setting up the model.

Suppose that a set of possible outcomes $(1, \dots, n)$, which cannot necessarily be observed directly, are such that:

$$P(\text{outcome } i) = \theta(i), \quad \sum_{i=1}^n \theta(i) = 1, \quad \theta(i) > 0, \quad 1 \leq i \leq n.$$

A sample $r = (r_1, \dots, r_m)$ of m possible records is observed where it will be assumed that $m \geq n$. Let Z denote the $n \times m$ matrix with z_{ij} as its $(i, j)^{th}$ component where

$$z_{ij} = \frac{P(\text{record } j \mid \text{outcome } i)}{P(\text{record } j \mid \text{outcome } i^*(j))} \quad (7.1)$$

where $i^*(j)$ is the outcome for which

$$\lambda_j = P(\text{record } j \mid \text{outcome } i^*(j)) = \max_{1 \leq i \leq n} P(\text{record } j \mid \text{outcome } i).$$

Thus, λ_j , $1 \leq j \leq m$, is the probability that brand j is chosen by a type of purchaser who likes it best. These probabilities are natural quantities of interest in this context since they are the brand purchasing probabilities associated with the customers targeted by competitive strategies. This definition implies that,

$$0 \leq z_{ij} \leq 1, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m. \quad (7.2)$$

Write $\theta = (\theta(1), \dots, \theta(n))^T$ and $\lambda = (\lambda_1, \dots, \lambda_m)^T$. Note that for non-degeneracy $0 < \lambda_j < 1$, $1 \leq j \leq m$, and henceforth this will be assumed. Now, $\lambda_j > 0$ implies

that brand j is bought by at least one type and so each column of Z has at least one component equal to 1 and $\lambda_j < 1$ implies that each customer buys at least two brands and so each row of Z must have at least two strictly positive terms.

For each outcome i , $1 \leq i \leq n$:

$$\sum_{j=1}^m P(\text{record } j \mid \text{outcome } i) = 1.$$

This can be rewritten to give the constraint:

$$\sum_{j=1}^m z_{ij} \lambda_j = 1, \quad 1 \leq i \leq n \quad (7.3)$$

It will be assumed that an observation vector \mathbf{r} of records gives rise to the likelihood:

$$L(\boldsymbol{\psi} \mid \mathbf{r}) = \prod_{j=1}^m \psi_j^{r_j}, \quad (7.4)$$

such that $\sum_{j=1}^m \psi_j = 1$ and $\psi_j > 0$, $1 \leq j \leq m$, where $N = \sum_{j=1}^m r_j$, $\boldsymbol{\psi}^T = (\psi_1, \dots, \psi_m)$ and $\psi_j = P(\text{record } j)$. For example, as for the Dirichlet model, $\{\mathbf{r} \mid \boldsymbol{\psi}\}$ could have a multinomial distribution. Since

$$\begin{aligned} \psi_j &= \sum_{i=1}^n P(\text{record } j \mid \text{outcome } i) \theta(i) \\ &= \lambda_j \sum_{i=1}^n z_{ij} \theta(i) \end{aligned} \quad (7.5)$$

then as Dickey et al (1987) point out, given the matrix Z , the likelihood $L(\boldsymbol{\psi} \mid \mathbf{r})$ separates in $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Thus inferences about $\boldsymbol{\psi}$ can be made directly from inferences about $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ through two likelihoods, $L_1(\boldsymbol{\theta})$ and $L_2(\boldsymbol{\lambda})$, such that:

$$L(\boldsymbol{\psi} \mid \mathbf{r}, Z) = L(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \mathbf{r}, Z) = L_1(\boldsymbol{\theta}) L_2(\boldsymbol{\lambda})$$

where

$$L_1(\boldsymbol{\theta}) = \prod_{j=1}^m (\xi_j(\boldsymbol{\theta}))^{r_j}, \quad \sum_{i=1}^n \theta(i) = 1, \quad 1 \leq i \leq n, \quad (7.6)$$

such that:

$$\xi_j(\theta) = \sum_{i=1}^n z_{ij}\theta(i), \quad 1 \leq j \leq m, \quad (7.7)$$

$$\text{i.e.} \quad \xi(\theta) = Z^T \theta, \quad \text{where} \quad \xi(\theta) = (\xi_1(\theta), \dots, \xi_m(\theta))^T$$

and where

$$L_2(\lambda) = \prod_{j=1}^m \lambda_j^{r_j}, \quad \sum_{j=1}^m z_{ij}\lambda_j = 1, \quad 1 \leq i \leq n \quad (7.8)$$

$$\text{i.e.} \quad Z\lambda = 1_n, \quad 0 < \lambda < 1_m$$

where, using the notation of the previous chapters, 1_q denotes a q -dimensional vector of 1's. It follows that given the prior independence of θ and λ , given Z , then θ and λ remain independent aposteriori. This property will be used throughout this chapter.

Using the condition that $m \geq n$ and any appropriate regularity conditions on λ , θ and Z , it looks likely that the solution space for λ , under the constraints of equation 7.8, is a manifold of dimension $m - n$ and the solution space for θ , under the constraint $\sum_{i=1}^n \theta_i = 1$, is of dimension $n - 1$. So it appears that Z has been used to reparameterise the $m - 1$ dimensional solution space for the vector of probabilities $\psi = (\psi_1, \dots, \psi_m)^T$ into an $m - 1 (= m - n + n - 1)$ dimensional manifold, the constrained solution space in (λ, θ) . However, unfortunately the necessary regularity conditions on Z to ensure this reparameterisation are non-trivial. Furthermore, it is simple to find matrices Z , all of whose components lie between zero and one, but for which constraints 7.8 can never be satisfied. Z will usually need to be specified directly as it forms the basis of the hypothesised model. It is therefore important to identify classes of matrices called *compatible matrices* which are matrices of likelihood ratios of the required form which satisfy all the constraints to allow the reparameterisation from ψ to (θ, λ) . This issue will be addressed in the next section.

Given a compatible matrix Z , a prior to posterior analysis of λ is possible

using a Dirichlet prior distribution (as given in equation 2.1 of chapter 2). However, before a prior to posterior analysis of θ is possible, another important issue concerning the form of Z is whether, under an appropriately chosen prior distribution, Bayes estimates of θ , as $N = \sum_{j=1}^m r_j \rightarrow \infty$, are consistent and whether certain combinations of θ are identifiable from τ . It is shown in Appendix A that when τ is multinomial, a necessary and sufficient condition for θ to be both consistent and identifiable, is that Z is a matrix of rank n — the number of outcome types. It will therefore be assumed throughout this chapter that Z is of rank n . Given a compatible Z of rank n , an exact prior to posterior analysis of θ is then possible using Generalised Dirichlet prior distributions (Dickey et al., 1987). The natural family of conjugate priors to use for θ is then given by:

$$p(\theta) \propto \prod_{j=1}^m \{\xi_j(\theta)\}^{\alpha_j}$$

where $\xi_j(\theta)$ is defined by equation 7.7 and $\alpha_j > 0$, for $1 \leq j \leq m$. The proportionality constant is a complicated function of Gamma functions which are tabulated as two-way multiple hypogeometric functions (Carlson, 1977, Dickey, 1983). This prior and the likelihood $L_1(\theta)$ enable posterior moments to be found explicitly. Posterior modes of θ are also easy to calculate since the posterior density is log concave.

7.3 Model Consistent Solutions.

In this section two issues are addressed. The first is the investigation of the form of compatible matrices, Z , so that a solution for λ exists when Z and λ satisfy the constraints 7.8. The second investigates how λ can be reparameterised while satisfying constraints 7.8 so that relatively simple prior to posterior analyses can be performed on these conditional probabilities. This section therefore defines a

class of Z matrices which not only admits consistent solutions but also has an $m - n$ dimensional manifold as a solution space which can be reparameterised in a simple way to give a straightforward prior to posterior analysis.

Call Z *recursive* if there is a sequence $\{k(i)\}_{1 \leq i \leq n}$ which is strictly increasing in i with

1. $z_{ij} \geq 0, \quad 1 \leq j \leq k(i) - 1$
2. $z_{i,k(i)} > 0$
3. $z_{ij} = 0, \quad k(i) < j \leq m, \quad k(n) = m.$

For example, the matrix Z is recursive where Z is given by:

$$Z = \begin{pmatrix} .5 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & .2 & 0 & 1 & 0 & 0 & 0 \\ .6 & .5 & 1 & .2 & 0 & 1 & 0 \\ 0 & 0 & .5 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Here $k(1) = 2, k(2) = 4, k(3) = 6$ and $k(4) = 7 = m$.

Call Z *recursive-directed* if it is recursive and when $j \leq k(i - 1)$, for every $2 \leq i \leq n$ there is a row $p(i)$, where $1 \leq p(i) \leq i - 1$ such that $z_{ij} \leq z_{p(i),j}$, with strict inequality for some $j, 1 \leq j \leq k(i - 1)$.

The matrix:

$$Z = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ .5 & .9 & 1 & 1 & 0 & 0 & 0 \\ .4 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

is recursive-directed. In this example $k(1) = 3, k(2) = 4, k(3) = 5$ and $k(4) = 7 = m$. Now,

$$\begin{aligned} p(2) &= 1 \quad \text{as} \quad z_{21} < z_{11}, \quad z_{22} < z_{12} \\ p(3) &= 1 \quad \text{as} \quad z_{31} < z_{11}, \quad z_{32} = z_{12}, \quad z_{33} < z_{13}, \quad z_{34} = z_{14} \\ p(4) &= 2 \quad \text{as} \quad z_{41} < z_{21}, \quad z_{42} < z_{22}, \quad z_{43} = z_{23}, \quad z_{44} = z_{24}, \quad z_{45} = z_{25} \end{aligned}$$

and

$$\begin{aligned} p(3) &\neq 2 & \text{as } z_{32} > z_{21} \\ p(4) &\neq 1 & \text{as } z_{44} > z_{14} \\ p(4) &\neq 3 & \text{as } z_{43} > z_{33} \end{aligned}$$

It will now be shown that when Z is recursive-directed, a solution for λ always exists which satisfies constraint 7.8 and a conjugate prior to posterior analysis is relatively simple to perform.

Write $\lambda^T = (\lambda(1)^T, \dots, \lambda(n)^T)$ where the component vector $\lambda(1)$ is labelled as follows,

$$\begin{aligned} \lambda(1)^T &= (\lambda_1, \dots, \lambda_{k(1)}), \\ &= (\lambda_1(1), \dots, \lambda_{t(1)}(1)), \\ \lambda(i)^T &= (\lambda_{k(i-1)+1}, \dots, \lambda_{k(i)}), \\ &= (\lambda_1(i), \dots, \lambda_{t(i)}(i)) \quad t(i) = k(i) - k(i-1), \quad 2 \leq i \leq n. \end{aligned}$$

Consider the segment of row i of Z :

$$z_i(l)^T = (z_{i,(l-1)+1}, \dots, z_{i,k(l)}), \quad 1 \leq l \leq i \leq n.$$

For example, suppose that:

$$Z = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0.1 & 1 & 1 \end{pmatrix} \quad (7.9)$$

then for example, $z_2(1)^T = (0, 0.5)$ and $z_3(3)^T = (1, 1)$. Note that since $\lambda_j > 0$, for $j = 1, \dots, m$, and Z is directed, implying that the lowest indexed row in which $z_{ij} > 0$ must be maximal over i , then:

$$z_l(l) = 1_{t(l)}, \quad 1 \leq l \leq n \quad (7.10)$$

where $1_{t(l)}$ is the row vector of $t(l)$ ones.

Let $x_i(l) = z_i(l)^T \lambda(l)$ such that, by the definition of recursive-directed, for any λ , $0 < \lambda_j < 1$, $1 \leq j \leq m$, there is a $p(i)$, $1 \leq p(i) < i$, for which:

$$x_i(l) \leq x_{p(i)}(l), \quad 1 \leq l \leq p(i), \quad 2 \leq i \leq n \quad (7.11)$$

with strict inequality for some $1 \leq l \leq p(i)$. For the Z matrix in equation 7.9, for example,

$$\begin{aligned} x_1(1) &= \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \\ x_2(1) &= \begin{pmatrix} 0 & 0.5 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \end{aligned}$$

and so $x_2(1) < x_1(1)$.

Define $\bar{x}_i(l) = z_i(l)^T \bar{\lambda}(l)$ where $\bar{\lambda}$ is a solution for λ which satisfies:

$$Z\bar{\lambda} = 1.$$

It will now be shown how this set of constraints on λ , when Z is recursive-directed, can be rewritten. The set of constraints exactly requires that there exists $\lambda(i) = \bar{\lambda}(i) > 0$, $1 \leq i \leq n - 1$ such that

$$\sum_{l=1}^i \bar{x}_i(l) = 1, \quad 1 \leq i \leq n. \quad (7.12)$$

Of particular interest are the linear combinations

$$\dot{\lambda}_l = \bar{x}_l(l) = z_l(l)^T \bar{\lambda}(l) \quad (7.13)$$

where by equation 7.10

$$\dot{\lambda}_l = \sum_{i=1}^{l(l)} \bar{\lambda}_i(l). \quad (7.14)$$

Lemma 7.3.1 *Suppose Z is recursive-directed and $\dot{\lambda}_l$ defined above. Then*

- i) for fixed Z , $\dot{\lambda}_l$ is a unique linear function of $\lambda^{l-1} = (\lambda(1), \dots, \lambda(l-1))$, for $l \geq 2$.*

ii) the function $\dot{\lambda}_l(\lambda^{l-1})$ lies between 0 and 1.

Proof: Go by induction on r , the index of components of λ for which the constraints 7.11 are satisfied. Suppose $\bar{\lambda}$ is a solution and Z is recursive-directed. Clearly (ii) is satisfied for $r = 1$ by the positivity of $\bar{\lambda}$ and (i) is trivially satisfied.

For $r \geq 2$

$$\sum_{l=1}^{r-1} z_r(l) \bar{\lambda}(l) < \sum_{l=1}^{p(r)} z_{p(r)}(l) \bar{\lambda}(l) = \sum_{l=1}^{p(r)} \bar{x}_{p(r)}(l) = 1 \quad (7.15)$$

by equations 7.11 and 7.12. Because of this strict inequality, the constraint $\sum_{l=1}^{p(r)} \bar{x}_{p(r)}(l) = 1$ can be written as

$$\dot{\lambda}_r = \bar{x}_r(r) = 1 - \sum_{l=1}^{r-1} z_r(l) \bar{\lambda}(l) > 0 \quad (7.16)$$

by the above. Note that $\dot{\lambda}_r$ is a linear function of λ^{r-1} and is strictly positive regardless of the value of λ^{r-1} because of the inequality in equation 7.15. The inductive step is therefore complete, so the lemma is proved.

A theorem is now presented which shows that a solution for λ always exists which satisfies constraint 7.8 when Z is recursive-directed.

Theorem 7.3.2 *If Z is recursive-directed then Z is compatible. Furthermore, the solution space $\bar{\lambda}$ satisfying the constraint is a manifold of dimension $(m - n)$.*

Proof: Lemma 7.3.1 showed that for each $l = 1, \dots, n$ $\dot{\lambda}_l$ is a function of λ^{l-1} where:

$$\dot{\lambda}_l = \sum_{j=1}^{t(l)} \bar{\lambda}_j(l) > 0.$$

Therefore, to prove the theorem, it is sufficient to identify those $\bar{\lambda} > 0$ consistent with the constraints

$$\dot{\lambda}_l > 0, \quad 1 \leq l \leq n.$$

First note that

$$\dot{\lambda}_1 = \sum_{j=1}^{t(1)} \bar{\lambda}_j(1) = 1.$$

Writing $\rho_j(1) = \bar{\lambda}_j(1)$ the solution space requires that $\sum_{j=1}^{t(1)} \rho_j(1) = 1$ and, to ensure the positivity of $\lambda(1)$,

$$\rho(1) = (\rho_1(1), \dots, \rho_{t(1)}(1))^T > \mathbf{0}.$$

The space defined by such ρ is clearly a manifold of dimension $t(1) - 1$. Set

$$\bar{\lambda}_j(l) = \dot{\lambda}_j(\lambda^{l-1})\rho_j(l), \quad 1 \leq j \leq t(l). \quad (7.17)$$

Then for each possible value of $\bar{\lambda}^{l-1}$ this defines a manifold of dimension $t(l) - 1$ and spans the solution space under the constraint

$$\sum_{j=1}^{t(l)} \rho_j(l) = 1 \quad \rho(l) = (\rho_1(l), \dots, \rho_{t(l)}(l))^T > 0.$$

As this holds for $1 \leq l \leq n$, it now follows that the solution space of λ is dimension

$$\sum_{l=1}^n (t(l) - 1) = m - n$$

as required.

By Appendix A it is possible to perform a prior to posterior analysis on θ using generalised Dirichlet distributions. Now, when the matrix Z is recursive-directed the above theorem implies that there is also a conjugate Bayesian analysis of the conditional probabilities λ and this will be the emphasis of much of the rest of the chapter.

It will now be shown how λ can be reparameterised in terms of $\rho(1), \dots, \rho(n)$ to provide a relatively simple prior to posterior analysis of λ . From equation 7.17:

$$\lambda(i) = \dot{\lambda}_i \{ \rho_1(i), \dots, \rho_{t(i)}(i) \}, \quad i = 1, \dots, n \quad (7.18)$$

where

$$\dot{\lambda}_i = 1 - \sum_{l=1}^{i-1} \dot{\lambda}_l z_i(l)^T \rho(l).$$

Notice that $\dot{\lambda}_i$ is a polynomial function of $\rho(1), \dots, \rho(i-1)$ of order no greater than $i-1$. Using equation 7.18, the likelihood of λ can be transformed to ρ in a reasonably straightforward manner. Thus, when $Z\lambda = 1_n$

$$L_2(\lambda) = \prod_{j=1}^m \lambda_j^{r_j} = \prod_{i=1}^n \left[\dot{\lambda}_i^{\dot{r}_i} \prod_{k=1}^{t(i)} (\rho_k(i))^{r_k(i)} \right] \quad (7.19)$$

such that $\sum_{k=1}^{t(i)} \rho_k(i) = 1$, $\rho_k(i) > 0$, $1 \leq k \leq t(i)$, $1 \leq i \leq n$ where $r_k(i) = r_j$ when $\lambda_k(i) = \lambda_j$ and $\dot{r}_i = \sum_{1 \leq k \leq t(i)} r_k(i)$.

A natural choice of conjugate prior to this likelihood, expressed as a distribution on $\{\rho(1), \dots, \rho(n)\}$ would be given by:

$$p_2(\rho) \propto \prod_{i=1}^n \left[\dot{\lambda}_i^{\dot{\alpha}_i} \prod_{k=1}^{t(i)} \{\rho_k(i)\}^{\alpha_k(i)} \right] \quad (7.20)$$

where ρ and $\dot{\lambda}$ are as in L_2 with $\dot{\alpha}_i, \alpha_k(i) > 0$, $1 \leq i \leq n$, $1 \leq k \leq t(i)$ satisfying the constraint

$$\dot{\alpha}_i = \sum_{1 \leq k \leq t(i)} \alpha_k(i).$$

The posterior density will clearly now take the same form as equation 7.20 with $\dot{\alpha}_i$ and $\alpha_k(i)$ replaced by $\dot{\alpha}_i + \dot{r}_i$ and $\alpha_k(i) + r_k(i)$ respectively. These densities will be called *nested generalised Dirichlets*. It will be shown in example 7.3.1 that their moments are straightforward to calculate for moderate sizes of N . For large N , since the posterior density is log concave, the posterior mode and its associated matrix of second derivatives of the log density are easy to calculate numerically. Often further simplifications to recognised structures are possible. Here are two examples.

7.3.1 Example.

Suppose that 5 brands are bought by 3 types of customer and the hypothesised model gives the likelihood ratio matrix Z given in equation 7.9. Note that the solution space for λ is spanned by $(\rho_1(1), \rho_2(1))$ of dim 1, $\rho_1(2) = 1$ of dim 0 and $(\rho_1(3), \rho_2(3))$ of dim 1. From equation 7.16:

$$\dot{\lambda}_1 = 1 \qquad \dot{r}_1 = r_1 + r_2 \quad (7.21)$$

$$\begin{aligned} \dot{\lambda}_2 &= 1 - 0.5\lambda_2 & \dot{r}_2 &= r_3 \\ &= 1 - 0.5\rho_2(1) & & \end{aligned} \quad (7.22)$$

and similarly

$$\begin{aligned} \dot{\lambda}_3 &= 1 - 0.1\lambda_3 & \dot{r}_3 &= r_4 + r_5 \\ &= 1 - 0.1(1 - 0.5\rho_2(1)) \\ &= 0.9 + 0.05\rho_2(1) & & \end{aligned} \quad (7.23)$$

The likelihood given in equation 7.19 now becomes, as a function of ρ :

$$\begin{aligned} L_2(\rho) &= (1 - 0.5\rho_2(1))^{r_3} (0.9 + 0.05\rho_2(1))^{r_4+r_5} \rho_1(1)^{r_1} \rho_2(1)^{r_2} \rho_1(3)^{r_4} \rho_2(3)^{r_5} \\ &= L_2(\rho(1))L_2(\rho(3)) \end{aligned}$$

where

$$L_2(\rho(1)) = (1 - 0.5\rho_2(1))^{r_3} (0.9 + 0.05\rho_2(1))^{r_4+r_5} \rho_1(1)^{r_1} \rho_2(1)^{r_2}$$

such that $\rho_1(1) + \rho_2(1) = 1$ and

$$L_2(\rho(3)) = \rho_1(3)^{r_4} (1 - \rho_1(3))^{r_5}.$$

Thus the likelihood separates in $\rho(1)$ and $\rho(3)$. The nested generalised Dirichlet defined above sets $\rho(1) \amalg \rho(3)$ apriori with $\rho(3)$ having a Beta distribution and

$\rho(1)$ having a generalised Beta distribution. A posteriori, $\rho(1)$ and $\rho(3)$ remain independent, $\rho(1) | \mathbf{r}$ having a generalised Beta density and $\rho(3) | \mathbf{r}$ having a Beta density with the usual updating equations relating posterior hyperparameters to prior hyperparameters and data \mathbf{r} .

To illustrate the Bayes updating, take Z as in equation 7.9 with $\mathbf{r}^T = (2, 4, 2, 1, 0)$ and $\rho(1)$ and $\rho(3)$ independent uniform priors. The moments of $\rho(1)$ posterior to \mathbf{r} can be calculated from tables of hypergeometric functions or more simply, in this case, by noting that the posterior density is a mixture of 4 Beta densities with mean 0.139 and variance 0.047. The posterior distribution of $\rho(3)$ is just Beta(2,1) and so has mean $\frac{2}{3}$ and variance $\frac{1}{18}$. Using equations 7.21, 7.22 and 7.23 it is clear that:

$$\begin{aligned}\lambda_1 &= \rho_1(1) \\ \lambda_2 &= 1 - \rho_1(1) \\ \lambda_3 &= 0.5(1 + \rho_1(1)) \\ \lambda_4 &= \{0.9 + 0.05(1 - \rho_1(1))\} \rho_1(3) \\ \lambda_5 &= \{0.9 + 0.05(1 - \rho_1(1))\} (1 - \rho_1(3)).\end{aligned}$$

Thus that the posterior mean and covariance of λ can be found and are given by:

$$E(\lambda^T | \mathbf{r}) = (0.139, 0.861, 0.5695, 0.6287, 0.314)$$

and

$$\text{cov}(\lambda | \mathbf{r}) = \begin{pmatrix} .0470 & -.0470 & .0240 & -.0016 & -.00075 \\ & .0470 & -.0235 & -.0613 & .0011 \\ & & .0117 & .1784 & -.0002 \\ & & & .0490 & -.0500 \\ & & & & .0494 \end{pmatrix}.$$

7.3.2 Example.

This time 7 brands are bought by 4 customers, so that the likelihood ratio matrix is:

$$Z = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & .5 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & .5 & 1 \end{bmatrix}$$

The solution space here is spanned by $(\rho_1(1), \rho_1(2))$ of dim 1, $(\rho_1(2), \rho_2(2))$ of dim 1, $(\rho_1(3), \rho_2(3))$ of dim 1 and $\rho_1(4) (= 1)$ of dim 0. This time

$$\dot{\lambda}_1 = 1$$

$$\dot{\lambda}_2 = 1 - \rho_2(1) = \rho_1(1)$$

$$\dot{\lambda}_3 = 1 - (0.5\rho_1(2) + \rho_2(2))\dot{\lambda}_2 = 1 - (1 - 0.5\rho_1(2))\rho_1(1)$$

$$\dot{\lambda}_4 = 1 - 0.5\rho_2(3)\dot{\lambda}_3$$

So $L_2(\rho)$ can be written as:

$$L_2(\rho) = L_2(\rho(1))L_2(\rho(2)|\rho(1))L_3(\rho(3)|\rho(1))$$

where

$$L_1(\rho(1)) = \rho_1(1)^{r_1+r_3+r_4}\rho_2(1)^{r_2} \quad \rho_1(1) + \rho_2(1) = 1$$

$$L_2(\rho(2)|\rho(1)) = \rho_1(2)^{r_3}\rho_2(2)^{r_4}\dot{\lambda}_3^{r_5+r_6} \quad \rho_1(2) + \rho_2(2) = 1$$

$$L_3(\rho(3)|\rho(1), \rho(2)) = \rho_1(3)^{r_5}\rho_2(3)^{r_6} \left\{1 - 0.5\rho_2(3)\dot{\lambda}_3\right\}^{r_7} \quad \rho_1(3) + \rho_2(3) = 1.$$

For simplicity assume that apriori $\rho_1(1)$, $\rho_1(2)$ and $\rho_1(3)$ have independent Beta distributions so that their joint density takes the product form $f_1(\rho(1))$, $f_2(\rho(2))$, $f_3(\rho(3))$. Then aposteriori, the posterior joint density of ρ takes the form,

$$f_3(\rho(3)|\rho(1), \rho(2), \tau) = I_3^{-1}L_3(\rho(3)|\rho(1), \rho(2))f_3(\rho(3))$$

$$f_2(\rho(2)|\rho(1), \tau) = I_2^{-1}I_3L_2(\rho(2)|\rho(1))f_2(\rho(2))$$

$$f_1(\rho(1)|\tau) = I_1^{-1}I_2L_1(\rho(1))f(\rho(1))$$

where I_3 , I_2 and I_1 are the proportionality constants that ensure $f_3(\cdot|\mathbf{r})$, $f_2(\cdot|\mathbf{r})$ and $f_1(\cdot|\mathbf{r})$ integrate to unity, I_2 and I_3 being functions of $\rho(1)$ and \mathbf{r} .

Note that $f_3(\rho(3)|\rho(1), \rho(2), \mathbf{r})$ is a generalised Beta density. Furthermore since λ_3 is a polynomial in $\rho(1)$ of degree 2, I_3 is a polynomial in $\rho(1)$ of degree $2r_7$, making f_2 a mixture of generalised Beta densities. Similarly I_2 is a polynomial of degree $2(r_5 + r_6 + r_7)$ so a posteriori $\rho(1)|\mathbf{r}$ is a mixture of generalised Betas. So in particular posterior moments can be calculated.

Bayesian prior to posterior analyses are even more straightforward if it is possible to reparametrise λ so that its likelihood is conjugate to a set of independent Dirichlet distributions.

7.4 The class of simple Z matrices.

One structure of Z which merits special attention is the class which satisfy the *differentially non-informedness hypothesis* (d.n.h.) in which the entries of Z are each either 0 or 1. In practical terms of looking at partially segmented markets, this means that each type of customer with a positive probability of buying a given brand, has an equal probability of buying that brand. Therefore, as it may be difficult to give conditional probabilities of records in reality, assuming that Z satisfies the d.n.h. can provide a good null hypothesis to work from.

Denote $C(i) = \{j : z_{ij} > 0\}$, for $1 \leq i \leq n$, and call $C(i)$ the *outcome clique* for outcome i . Call a model *tree-like* if Z is d.n.h. recursive-directed with the additional property that

- i) $z_l(l) = 1$ is one dimensional $2 \leq l \leq n$
- ii) each row has the same number of elements, $m^* = (m - n + 1)$, so that each outcome clique contains the same number of possible records.

By theorem 7.3.2, the solution space is of dimension $m-n$ and is parameterised by

$$\lambda(1) = (\lambda_1(1), \dots, \lambda_k(1))^T, \quad \sum_{i=1}^k \lambda_i(1) = 1, \quad k = k(1) = m - n + 1.$$

This is because i) and ii) then demand that the scalar $\dot{\lambda}_l$, $2 \leq l \leq n$, satisfies:

$$\dot{\lambda}_l = \lambda_j(1) \quad \text{for some } 1 \leq j \leq k. \quad (7.24)$$

For example, let:

$$Z = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

then

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

therefore:

$$\lambda_4 = 1 - (\lambda_1 + \lambda_3) = \lambda_2$$

and

$$\lambda_5 = 1 - (\lambda_3 + \lambda_4) = \lambda_1$$

since from row 2 it is known that $\lambda_1 + \lambda_3 + \lambda_4 = 1$.

Let $\bar{r}_j = r_j + \sum_{i \in A_j} r_i$ where $A_j = \{l : \lambda_j(1) = \dot{\lambda}_l, 2 \leq l \leq n\}$ for $1 \leq j \leq k$.

Then

$$L_2(\lambda) = \prod_{j=1}^k \lambda_j(1)^{\bar{r}_j}, \quad \sum_{j=1}^k \lambda_j(1) = 1$$

with the $n-1$ constraints 7.24 relating $\dot{\lambda}_l$, $2 \leq l \leq n$ to a $\lambda_j(1)$. Obviously a Dirichlet, $Di(\alpha)$, $\alpha^T = (\alpha_1, \dots, \alpha_k)$ on $\lambda(1)$ (and hence λ degenerately) is conjugate to this likelihood and the posterior of $\lambda(1) | r$ is $Di(\alpha^*)$, $\alpha^{*T} = (\alpha_1^*, \dots, \alpha_k^*)$ where $\alpha_j^* = \alpha_j + \bar{r}_j$. The constraints 7.24 now give the full posterior distribution on λ . This class can be further generalised to give a product Dirichlet form.

Call Z *simple* if it can be expressed as

$$Z = Z^{(1)} Z^{(2)}$$

where $Z^{(1)}$ is an $n \times m'$ tree-like structure and $Z^{(2)}$ is an $m' \times m$ partitioning matrix which by definition has exactly one non-zero term in each column and at least one non-zero term in each row. The study of Bayesian inference under Z of the form $Z^{(2)}$ is studied in some depth in Dickey et al (1987).

Without loss it can be assumed that $Z^{(2)}$ is recursive-directed by relabelling the components of λ so that $Z^{(2)}$ is a block diagonal matrix with a row vector of $t^{(2)}(i)$ 1's on its i^{th} diagonal. This means that if $z_{ij}^{(2)}$ denotes the $(i, j)^{th}$ element of $Z^{(2)}$, then for each row i :

$$\begin{aligned} z_{ij}^{(2)} &= 1 \quad \text{for } k^{(2)}(i-1) + 1 \leq j \leq k^{(2)}(i) \\ &= 0 \quad \text{otherwise} \end{aligned}$$

such that $\{k^{(2)}(i)\}_{1 \leq i \leq m'}$ is a strictly increasing sequence in i with $k^{(2)}(m') = m$.

For example, the matrix Z is simple, where:

$$Z = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

such that:

$$Z^{(1)} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad Z^{(2)} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

In this case, $m' = 3$ and, using the notation of section 7.3, Z has values $k(1) = 5$, $k(2) = 7$, $t(1) = t(2) = 5$, whereas $k^{(2)}(1) = 2$, $k^{(2)}(2) = 5$, $k^{(2)}(3) = 7$ and $t^{(2)}(1) = 2$, $t^{(2)}(2) = 3$, $t^{(2)}(3) = 2$. Notice that $\sum_{i=1}^{m'} t^{(2)}(i) = m$ for the partitioning matrix $Z^{(2)}$.

Partition the vector λ so that:

$$\lambda^T = (\bar{\lambda}(1), \dots, \bar{\lambda}(m'))$$

where

$$\begin{aligned}\bar{\lambda}(i)^T &= (\lambda_{k^{(2)}(i-1)+1}, \dots, \lambda_{k^{(2)}(i)}) \\ &= (\lambda_1(i), \dots, \lambda_{t^{(2)}(i)}(i))\end{aligned}$$

Let $\tau = Z^{(2)}\lambda$, where $\tau^T = (\tau_1, \dots, \tau_{m'})$ so that

$$\tau_i = \sum_{k=1}^{t^{(2)}(i)} \lambda_k(i), \quad 1 \leq i \leq m'.$$

Theorem 7.3.2 now allows the $m - n$ dimensional solution space to be written in the form:

$$\lambda_k(i) = \tau_i \rho_k(i) \quad \sum_{k=1}^{t^{(2)}(i)} \rho_k(i) = 1, \quad \rho_k(i) > 0$$

where $1 \leq k \leq t^{(2)}(i)$, $1 \leq i \leq m'$, $\sum_{i=1}^{m'} t^{(2)}(i) = m$. Notice that the reparameterisation is dictated by the form of the partitioning matrix $Z^{(2)}$ rather than Z itself, as was previously the case.

Relabel the components of τ as $(\bar{\tau}_1, \dots, \bar{\tau}_{m'})$ where $\bar{\tau}_i^T = (r_1(i), \dots, r_{t(i)}(i))$ so that $r_j = r_k(i)$ whenever $\lambda_j = \lambda_k(i)$. Then $L_2(\lambda | \tau)$ can be written in a similar fashion to equation 7.19 so that:

$$L_2(\lambda | \tau) = \prod_{j=1}^m \lambda_j^{r_j} = \prod_{i=1}^{m'} \left[\tau_i^{\hat{r}_i} \prod_{k=1}^{t^{(2)}(i)} \rho_k(i)^{r_k(i)} \right]$$

where $\hat{r}_i = \sum_{k=1}^{t(i)} r_i(k)$. Since $Z^{(1)}$ is tree-like, a conjugate analysis of λ can be performed with independent Dirichlet priors on $\tau^T = \{\tau_1, \dots, \tau_{m'}\}$ and all the non-degenerate vectors $\rho(i)^T = (\rho_1(i), \dots, \rho_{t(i)}(i))$, $1 \leq i \leq m'$.

The next example shows how a prior to posterior analysis can be performed on a partially segmented market with a simple Z , while simultaneously incorporating a time series model on the data.

	TIME																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Sum
r_1	2	5	4	2	5	3	5	1	4	3	3	2	4	5	2	5	55
r_2	1	3	5	2	3	1	2	0	3	3	2	3	3	1	1	2	35
r_3	4	2	4	6	1	3	4	4	5	5	5	4	2	6	2	5	62
r_4	1	2	2	2	1	0	0	2	4	1	2	1	1	1	2	1	23
r_5	4	6	2	4	3	2	2	10	5	7	8	6	6	4	4	6	79
r_6	4	3	3	5	2	4	5	2	2	2	4	1	2	2	1	2	44
r_7	2	0	3	4	2	1	0	3	3	3	4	0	3	2	4	2	36

Table 7.1: Bi-weekly sales of 7 brands of computer (confidential industrial source).

7.4.1 Example.

Suppose that in a certain market there are 7 brands of computers which are believed to be purchased by 4 different types of customer. For the sake of illustration it will be assumed that the d.n.h. holds and that the Z matrix, is simple, where Z is given by:

$$Z = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

such that:

$$Z^{(1)} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad Z^{(2)} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

A multivariate time series of the bi-weekly sales of these 7 brands over 16 time periods is presented in table 7.1. It is assumed that this process continues to respect the conditional independence structure assumed in the model.

It is believed that over the 32 weeks of sales considered, the proportion of customers of different types θ remained unchanged. However, because of repricing and advertising effects, the probability vector λ was volatile.

At each time point the likelihoods on θ and λ given the vector of sales r are

respectively

$$L_1(\theta) = (\theta(1) + \theta(3))^{r_1+r_2} (\theta(1) + \theta(2))^{r_3} \theta(2)^{r_4} (\theta(3) + \theta(4))^{r_5+r_6} \theta(4)^{r_7}$$

by equation 7.6 and

$$L(\lambda) = \rho(1)^{r_1} (1 - \rho(1))^{r_2} \rho(3)^{r_3} (1 - \rho(3))^{r_6} \tau^{\{r_1+r_2+r_4+r_7\}} (1 - \tau)^{\{r_3+r_5+r_6\}}$$

where

$$\left. \begin{array}{ll} \lambda_1 = \rho(1)\tau & \lambda_5 = \rho(3)(1 - \tau) \\ \lambda_2 = (1 - \rho(1))\tau & \lambda_6 = (1 - \rho(3))(1 - \tau) \\ \lambda_3 = 1 - \tau & \lambda_7 = \tau \\ \lambda_4 = \tau & \end{array} \right\} \quad (7.25)$$

If a random variable X follows a Beta distribution, then the density for X is given by:

$$p(x | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1} \quad \alpha, \beta > 0 \quad 0 \leq x \leq 1.$$

Then set apriori $\rho(1)$, $\rho(2)$ and τ independent with

$$\begin{aligned} \{\rho(i) | D_0\} &\sim Be(\alpha_0(i), \beta_0(i)) \quad i = 1, 3 \\ \{\tau | D_0\} &\sim Be(\alpha_0(\tau), \beta_0(\tau)) \end{aligned}$$

The processes $\rho(1)$, $\rho(3)$ and τ remain independent over time and so using the simple power steady model of Smith (1979, 1990), the beliefs about $\rho(1)$, $\rho(3)$ and τ can be updated after observations have been made in a straightforward fashion. Using the notation of chapter 3, the posterior distributions of $\rho(1)$, $\rho(2)$ and τ , given the first t observations of the series $\mathbf{r}^t = \{\mathbf{r}_1^t, \dots, \mathbf{r}_7^t\}^T$, is given by:

$$\begin{aligned} \{\rho(i) | \mathbf{r}^t\} &\sim Be(\alpha_t(i), \beta_t(i)) \quad i = 1, 3 \\ \{\tau | \mathbf{r}^t\} &\sim Be(\alpha_t(\tau), \beta_t(\tau)) \end{aligned}$$

where the parameters of these distributions can be calculated recursively from

$$(\alpha_t(1) - 1) = 0.7(\alpha_{t-1}(1) - 1) + r_{1,t}$$

$$(\beta_t(1) - 1) = 0.7(\beta_{t-1}(1) - 1) + r_{2,t}$$

$$(\alpha_t(3) - 1) = 0.7(\alpha_{t-1}(3) - 1) + r_{5,t}$$

$$(\beta_t(3) - 1) = 0.7(\beta_{t-1}(3) - 1) + r_{6,t}$$

$$(\alpha_t(\tau) - 1) = 0.9(\alpha_{t-1}(\tau) - 1) + (r_1 + r_2 + r_4 + r_7)_t$$

$$(\beta_t(\tau) - 1) = 0.9(\beta_{t-1}(\tau) - 1) + (r_3 + r_5 + r_6)_t$$

The moments of $\rho(1)$, $\rho(3)$ and τ conditional on the past are now easily calculated. It is then simple, using equation 7.25, to find the moments of λ from the moments of $(\rho(1), \rho(3), \tau)$. Figure 7.1 plots the evolution of the means of the distributions representing the beliefs about $(\lambda_1 \dots \lambda_7)^T$ against time.

An interesting feature of this data set is that there appears to be a dramatic increase in sales of brand 5 after time point 8. Subsequent investigations concluded that this was due to an aggressive repricing. Note how the model quickly adapts to this feature, adjusting down its main competitor (brand 6) whilst not interfering with the evolution of other brands significantly. Of course, if prior knowledge of such repricing were available then the usual Bayes intervention procedures of section 3.4 would apply. The simplest way to intervene would be to increase $\alpha_8(3)/\alpha_8(3) + \beta_8(3)$ and increase the variance of $\rho(3)$ by setting $\alpha_8(3) + \beta_8(3)$ to a smaller value.

The analysis of the component θ is more straightforward as it was assumed that θ remained static with time. With a generalised Dirichlet prior on θ of the form suggested by Dickey et al. (1987), set the prior density for θ as:

$$p(\theta) \propto \{(\theta(1) + \theta(3))^{10} \theta(2)^2 \theta(4)^5\} \{(\theta(1) + \theta(2))^8 (\theta(3) + \theta(4))^{12}\}.$$

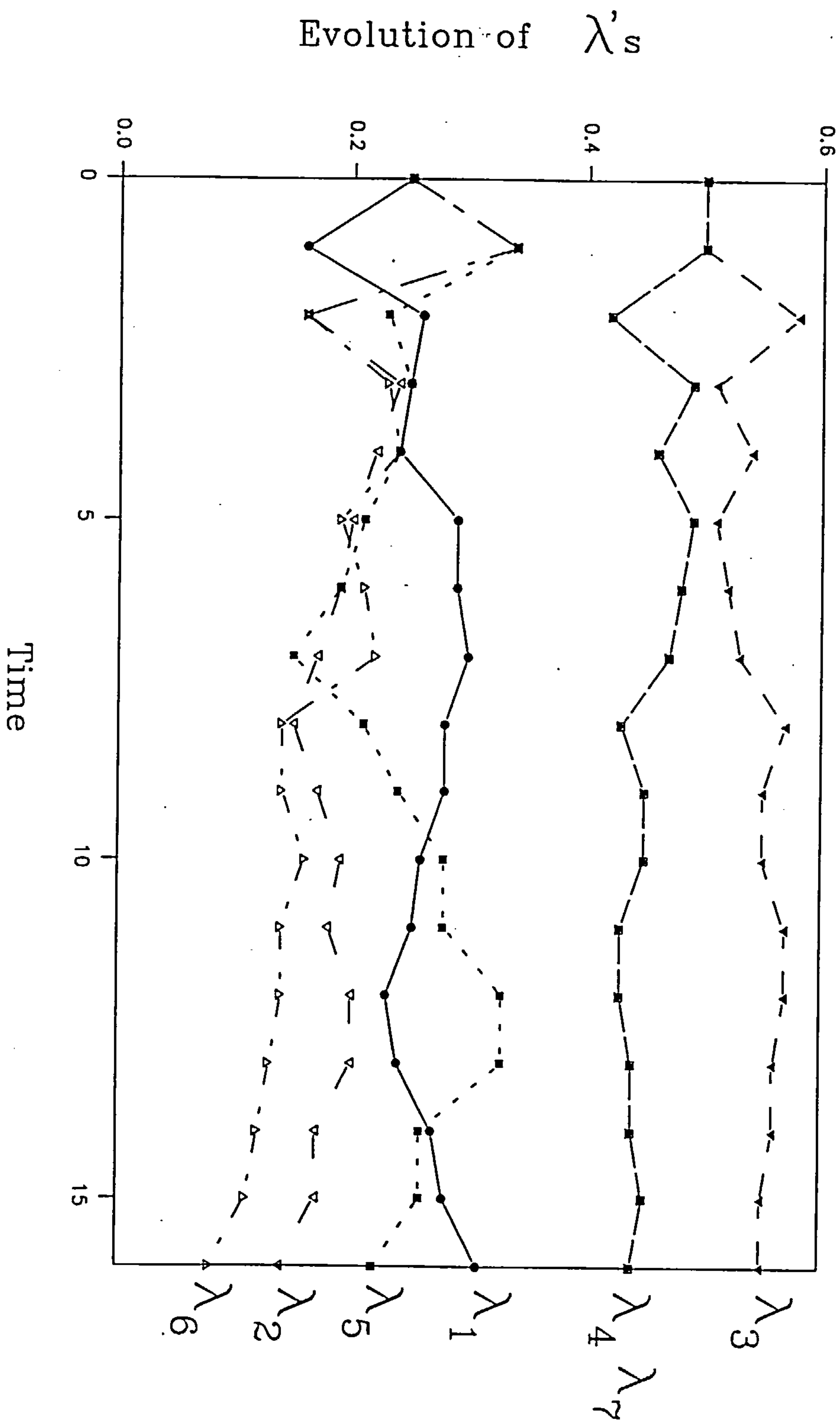


Figure 7.1: Time series of the evolution of the λ parameters.

After the 16 weeks, the posterior density is then of the form:

$$p(\theta|\mathbf{r}) \propto \{(\theta(1) + \theta(3))^{\gamma_1} \theta(2)^{\gamma_2} \theta(4)^{\gamma_3}\} \{(\theta(1) + \theta(2))^{\gamma_4} (\theta(3) + \theta(4))^{\gamma_5}\}$$

where $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (100, 25, 41, 70, 135)$. It is easily checked that the posterior mode $\hat{\theta} = (\hat{\theta}(1), \hat{\theta}(2), \hat{\theta}(3), \hat{\theta}(4))$ of θ is given by $= (0.19, 0.15, 0.41, 0.25)$. Since $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ are large the posterior joint density of θ is approximately normal about its mode with approximate covariance obtained by putting these estimates into the Fisher information matrix.

7.5 Representing Outcomes as Cliques on a Graph of Records.

It has been illustrated that it is technically possible to perform a variety of conjugate analyses on the class of problems so far discussed. However, because of the heterogeneity of this class it is often difficult to visualise the updating mechanism and the notation quickly becomes unwieldy. For a large class of Z matrices it is possible to represent its multivariate structure by a unique and illuminating graph. This is the subject of this section.

Recall that an outcome set is denoted by $C(k) = \{j : z_{kj} > 0\}$, $1 \leq k \leq n$ and let $Z^* = \{z_{ij}^*\}$ be the $n \times m$ matrix defined from Z by

$$\begin{aligned} z_{ij}^* &= z_{ij} & z_{ij} &= 0 \\ z_{ij}^* &= 1 & z_{ij} &> 0 \end{aligned} \quad 1 \leq i \leq n, \quad 1 \leq j \leq m.$$

Note that under the d.n.h. $Z^* = Z$.

The graph $G(Z^*)$ has m nodes representing the records on Z . Two record nodes j_1 and j_2 are connected by an (undirected) edge if and only if there exists an outcome set $C(i)$, $1 \leq i \leq n$ containing both nodes j_1 and j_2 . Call $G(Z^*)$ the *graph* of Z^* . Dechter & Pearl (1987) and Dechter et al (1990) call this the primal-constraint graph associated with the set of constraints imposed on

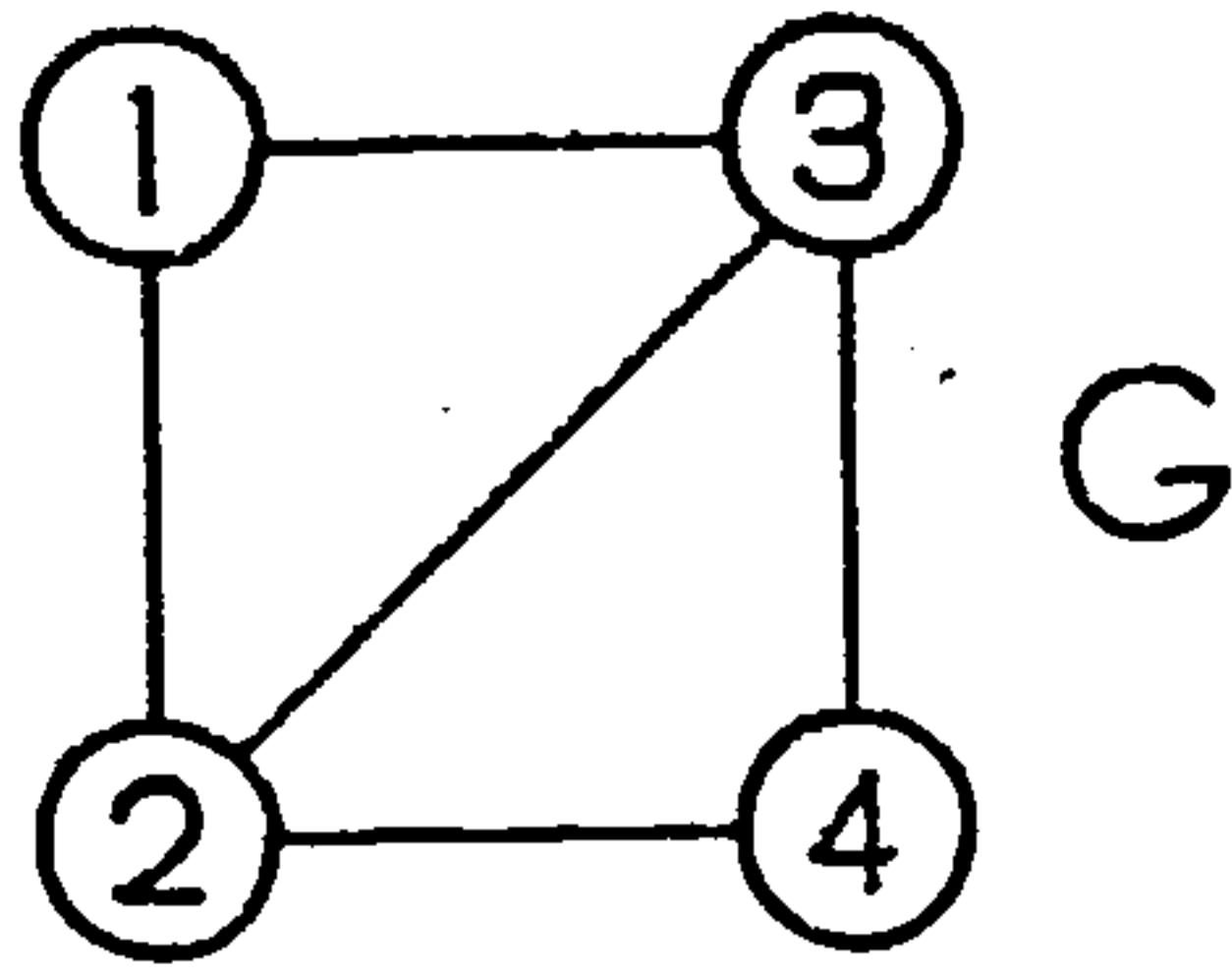


Figure 7.2: Graph G whose G -set includes the Z^* matrices given in equation 7.26.

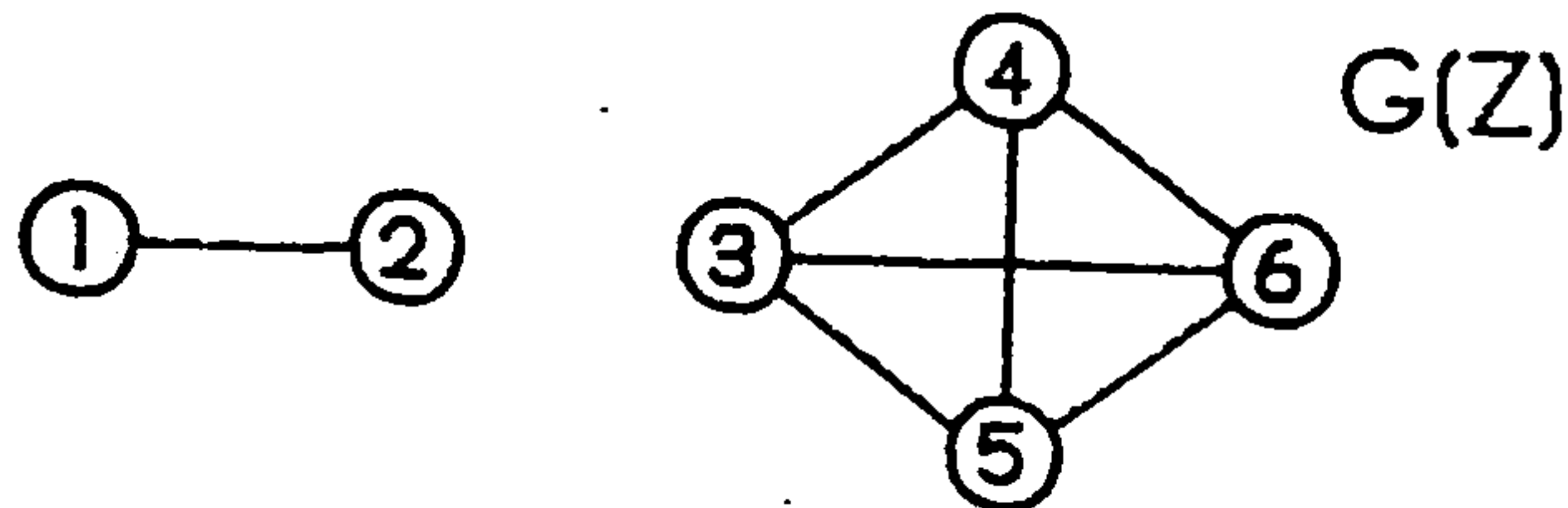


Figure 7.3: Graph $G(Z)$ of a partitioning matrix. Notice that $C(1)=\{1,2\}$ and $C(2)=\{3,4,5,6\}$.

λ . Clearly $G(Z^*)$ is well defined for each Z^* and the outcome sets $C(k)$, for $1 \leq k \leq n$ correspond to complete subgraphs of $G(Z^*)$ since all nodes in $C(k)$ are joined together. In general, for any graph G there is a set of Z^* , called the G -set, whose graph is G . For example, graph G_1 of figure 7.2 has a G -set which includes:

$$Z_1^* = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad Z_2^* = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} \quad (7.26)$$

and many other possibilities. If its outcome sets correspond to the cliques (i.e. maximal connected subgraphs) of G , then Z^* is called *graphical*. Dickey et al. (1987) considered the updating of models for graphs of partitioning matrices. The graph of a partitioning matrix is given in figure 7.3.

Call Z^* *abundant* when there is a subset of at least two rows (outcome sets) I , where each pair (j_1, j_2) of elements in $\cup_{i \in I} C(i)$ are such that $j_1, j_2 \in C(i^*)$, $i^* \in I$, or in other words, each pair (j_1, j_2) of elements in $\cup_{i \in I} C(i)$ can be found

in a row $i^* \in I$ of Z^* such that:

$$z_{i^*,j_1}^* = z_{i^*,j_2}^* = 1$$

For example, Z_2^* of equation 7.26 is abundant since all pairs (j_1, j_2) for $j_1, j_2 = 1, 2, 3$ appear together in one of the rows of $I = \{1, 2, 3\}$.

If Z^* is not abundant, then the complete subgraphs of $G(Z^*)$ formed from each outcome set cannot be combined to form another complete subgraph. (Subgraphs G_i , $i \in I$ are *combined* by taking a union of their nodes and edge sets). It follows that if Z^* is not abundant it must be graphical since each outcome set forms a clique, and if it is graphical it cannot be abundant by the maximality of cliques.

For most of the rest of this chapter it will be assumed that Z^* is graphical. This is a non-trivial restriction. For example, if some rows k_1 and k_2 of Z^* are nested, i.e.

$$C(k_1) \subseteq C(k_2) \quad 1 \leq k_1, k_2 \leq n$$

setting $I = \{k_1, k_2\}$ and noting $\bigcup_{i \in I} C(i) = C(k_2)$ it is obvious that Z^* is abundant and hence not graphical. Note, however, that nesting is precluded under the d.n.h. The reason for this is as follows. Since $Z = Z^*$ is of rank n , then any nesting must be of the form $C(k_1) \subset C(k_2)$. It follows that there is a component λ_j of λ such that $\lambda_j \in C(k_2) \setminus C(k_1)$ which by constraint 7.8 implies that $\lambda_j = 0$, contrary to the initial conditions imposed on λ . Less trivially, since any complete graph of more than 3 nodes can be expressed as a combination (for definition, see above) of at least 3 complete subgraphs, it is possible to construct Z which satisfy both the d.n.h and are abundant. For example, suppose that $Z(= Z^*) = \{z_{ij}\}$ is defined by

$$z_{1j} = \begin{cases} 1 & 1 \leq j \leq m-1 \\ 0 & \text{otherwise} \end{cases}$$

$$z_{2,j} = \begin{cases} 1 & 1 \leq j \neq m-1 \leq m \\ 0 & \text{otherwise} \end{cases}$$

$$z_{3,j} = \begin{cases} 1 & j = m-1 \text{ or } m \\ 0 & \text{otherwise} \end{cases}$$

In this case Z^* satisfies the d.n.h. and is also abundant. Clearly its graph is complete on m nodes although Z expresses the graph as a combination of 3 complete subgraphs. The graphical Z^* associated with $G(Z)$ simply consists of a single outcome set (row) on these m records.

Despite the restriction that Z^* is graphical, it will be shown later that Z^* within the same G-set exhibit very similar conditional independence structures on the variables of interest.

Recall from section 4.3 that a graph is called decomposable if it has the *running intersection property* (RIP), i.e. if its cliques can be indexed $C(1), \dots, C(n)$ so that

$$S(l) \left(= C(l) \cap \bigcup_{i=1}^{l-1} C(i) \right) \subseteq C(p(l)) \quad (7.27)$$

for some $p(l)$, $1 \leq p(l) \leq l-1$ this being true for $l = 2, \dots, n$.

The following result indicates that many useful models have Z^* which are graphical.

Theorem 7.5.1 *If Z (and hence Z^*) is recursive and has no nested outcome sets, i.e. no sets $C(i)$, $C(k)$ of non-zero elements of Z , $1 \leq i, k \leq n$ such that $C(i) \subseteq C(k)$, then*

i) Z^ is graphical*

ii) the graph G of Z^ is decomposable if in addition Z^* is recursive-directed.*

Proof: Suppose Z^* described above is abundant. Then by definition, in particular, there must exist an index set $I = \{i_1, \dots, i_k\}$, $i_1 < \dots < i_k$ of indices of

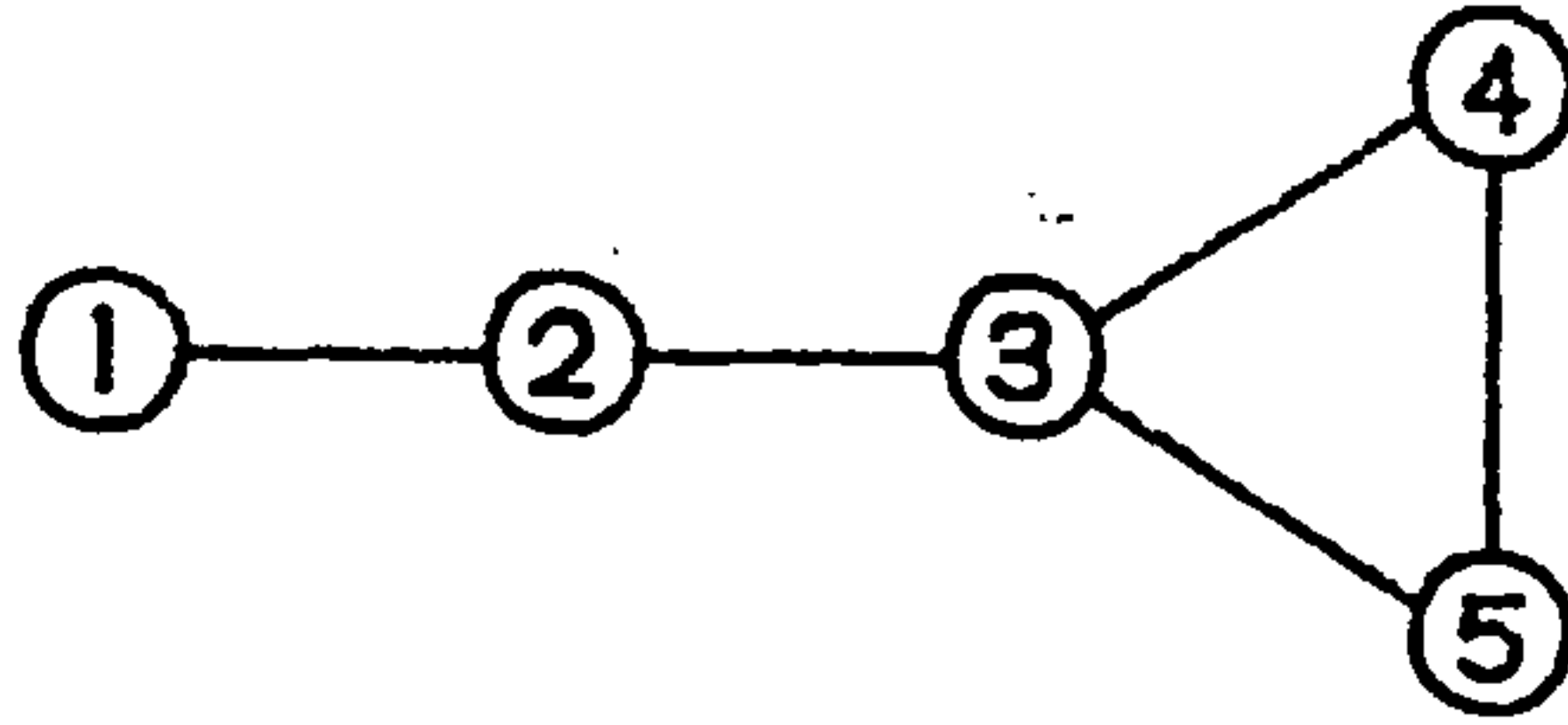


Figure 7.4: Graph of example 7.3.1.

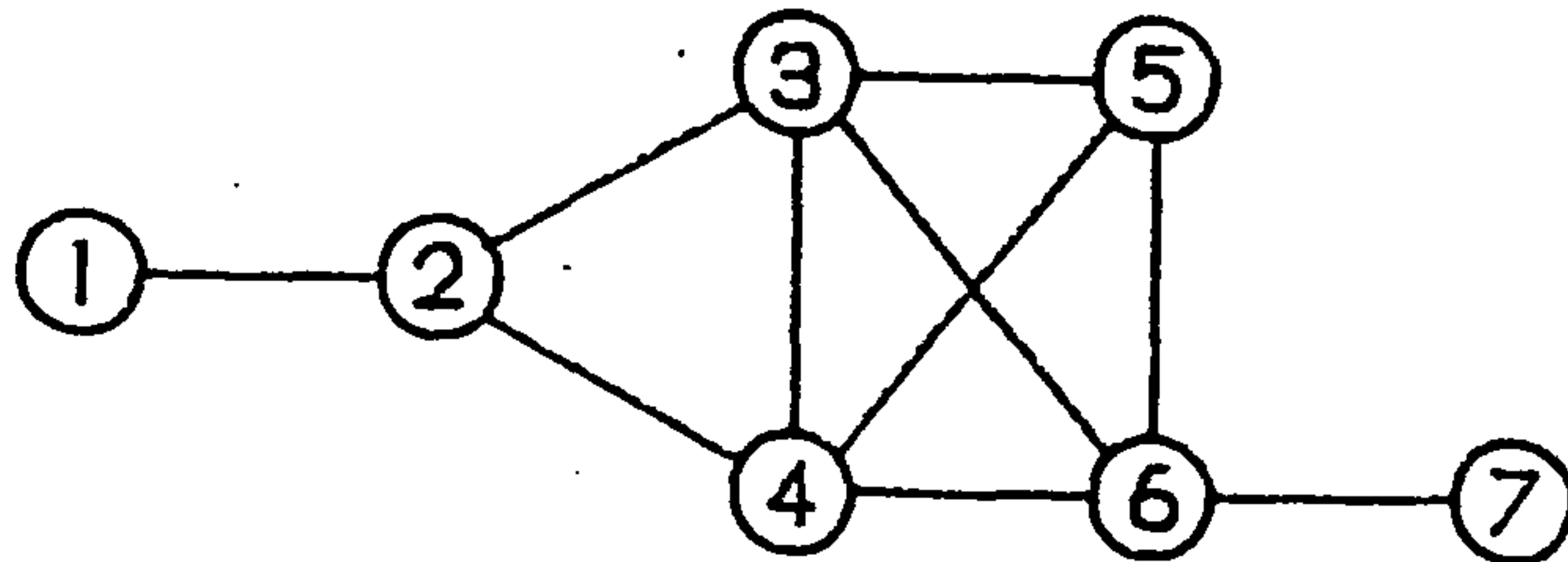


Figure 7.5: Graph of example 7.3.2.

the rows of Z^* whose associated cliques formed from these rows combine to form another clique. Let $j_k \in D(i_k) = C(i_k) \setminus \bigcup_{i \in I \setminus \{i_k\}} C(i)$. Because Z is recursive, $D(i_k)$ is non-empty and so such a j_k can be found. Since j_k lies only in $C(i_k)$ for $\bigcup_{i \in I} C(i)$ to have a complete subgraph, it follows that it must be true that for all $j \in C(i_1)$:

$$\{j, j_k\} \in C(i_k).$$

But this implies that

$$C(i_1) \subset C(i_k)$$

contrary to the hypothesis. So Z^* cannot be abundant and therefore must be graphical.

ii) Since Z^* satisfies the d.n.h. the assertion follows directly from equation 7.27.

The decomposable graphs of figures 7.4, 7.5 and 7.6 correspond to the recursive-directed Z^* of to examples 7.3.1, 7.3.2 and 7.4.1 respectively.

The graph in figure 7.7 contains 2 graphs G_1 and G_2 for which neither model is identifiable for θ , even though the d.n.h. model for G_1 has Z which is compatible

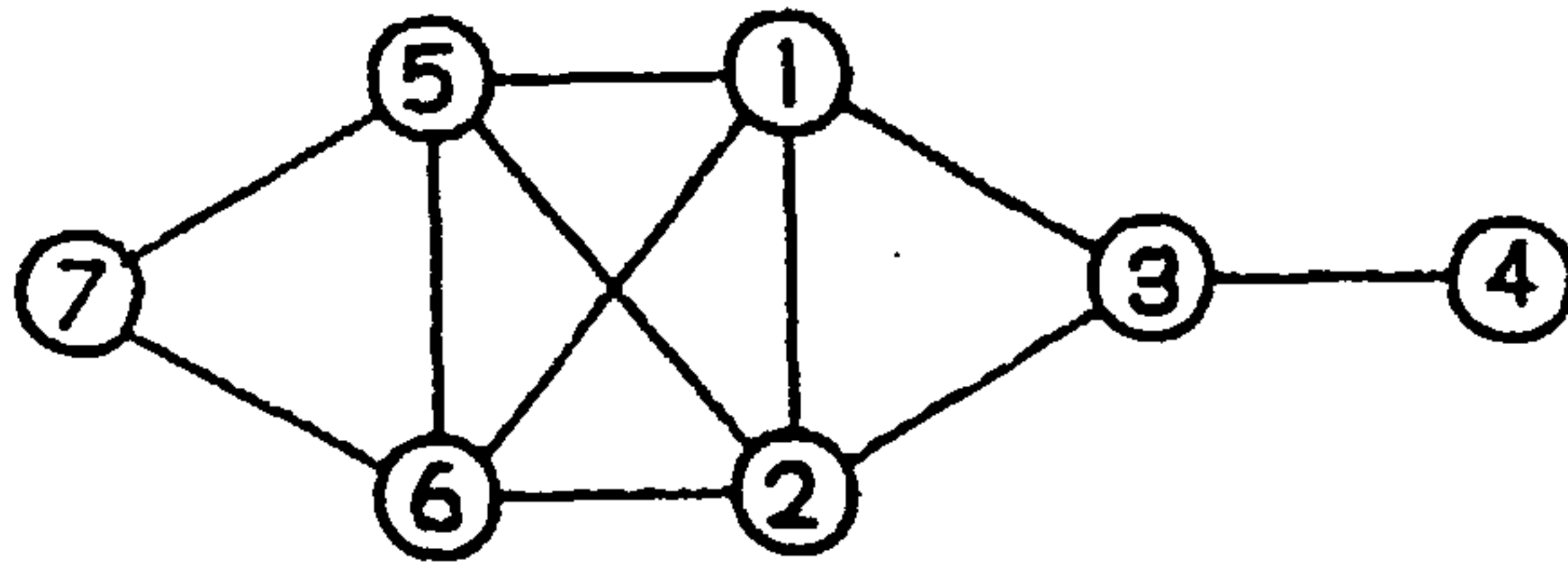


Figure 7.6: Graph of example 7.4.1.

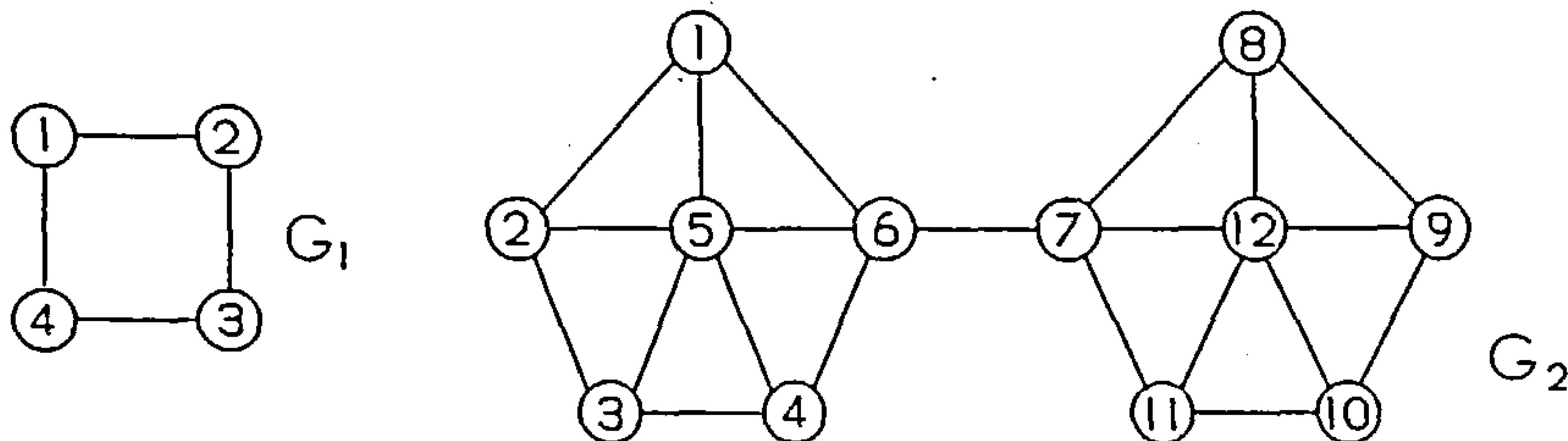


Figure 7.7: Two graphs which are non-identifiable for θ .

and at first sight it appears that for G_2 inferences about θ can be made using a generalised Dirichlet prior.

When a tree-like model has $m^* = 2$, then its graph G is a tree (hence the name). For general C , a tree-like model has $m - 1$ branches where each “branch” of $l - 1$ edges connects junctions of $l - 1$ nodes.

The motivation for representing Z by $G(Z^*)$ is straightforward. Obviously since $\sum_{j \in C(i)} z_{ij} \lambda_j = 1$

$$\lambda_k \Pi \lambda | n_i(\lambda_k), \quad \lambda_k \in C(i)$$

where $n_i(\lambda_k) = \{\lambda_j : z_{ij} > 0 \text{ and } z_{ik} > 0, j \neq k\} = \{\lambda_j : j, k \in C(i), j \neq k\}$. This in particular implies that

$$\lambda_k \Pi \lambda | n(\lambda_k) \tag{7.28}$$

where

$$n(\lambda_k) = \bigcup_{i=1}^n n_i(\lambda_k) = \{\lambda_j : j \text{ is connected to } k \text{ in } G(Z^*)\}.$$

So in particular, $G(Z^*)$ is the usual undirected graph representing the c.i. structure across λ . This property is conjugate in the sense that it is obviously pre-

served after observing \mathbf{r} . Note that these properties hold even when Z is not graphical. Statements 7.28 remain true for all Z^* in the corresponding G-set, so that Z with the same graph have similar conditional independence structures across λ .

7.6 D.n.h. Reparametrisation Using Graphical Results.

In Section 7.4 it was shown how Z matrices which are simple admit a conjugate product Dirichlet prior to posterior analysis for the conditional probabilities λ . In fact this parametrisation is invariant (up to indexing of $\rho(1) \dots \rho(n)$) to the choice of any Z within the G-set of G . However, under the d.n.h. two matrices Z and Z' within the G-set of G may both have an ordering of cliques which satisfies the RIP, implying each is recursive-directed, yet for which the reparametrisations of λ , as directed by theorem 7.3.2, are not equivalent. In particular one may admit a product Dirichlet form whilst the other does not.

For example, consider the 2 recursive-directed Z matrices

$$Z = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \text{ and } Z' = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad (7.29)$$

where $G(Z)$ and $G(Z')$ are given in figure 7.8. Clearly $G(Z)$ and $G(Z')$ are identical except for the labelling and as such represent the same partial segmentation. However under theorem 7.3.2 the two Z matrices give different parameterisations. The first gives a likelihood reparametrisation of the form

$$L_2(\rho) = \rho_1(1)^{r_1+r_4} \rho_2(1)^{r_2} \rho_3(1)^{r_3} (1 - \rho_2(1) + \rho_1(3))^{r_5}$$

and the second

$$L'_2(\rho') = \rho'_1(1)^{r_1+r_3+r_4+r_5} \rho'_2(1)^{r_2} \rho'_1(2)^{r_3+r_4} \rho'_2(2)^{r_3}.$$

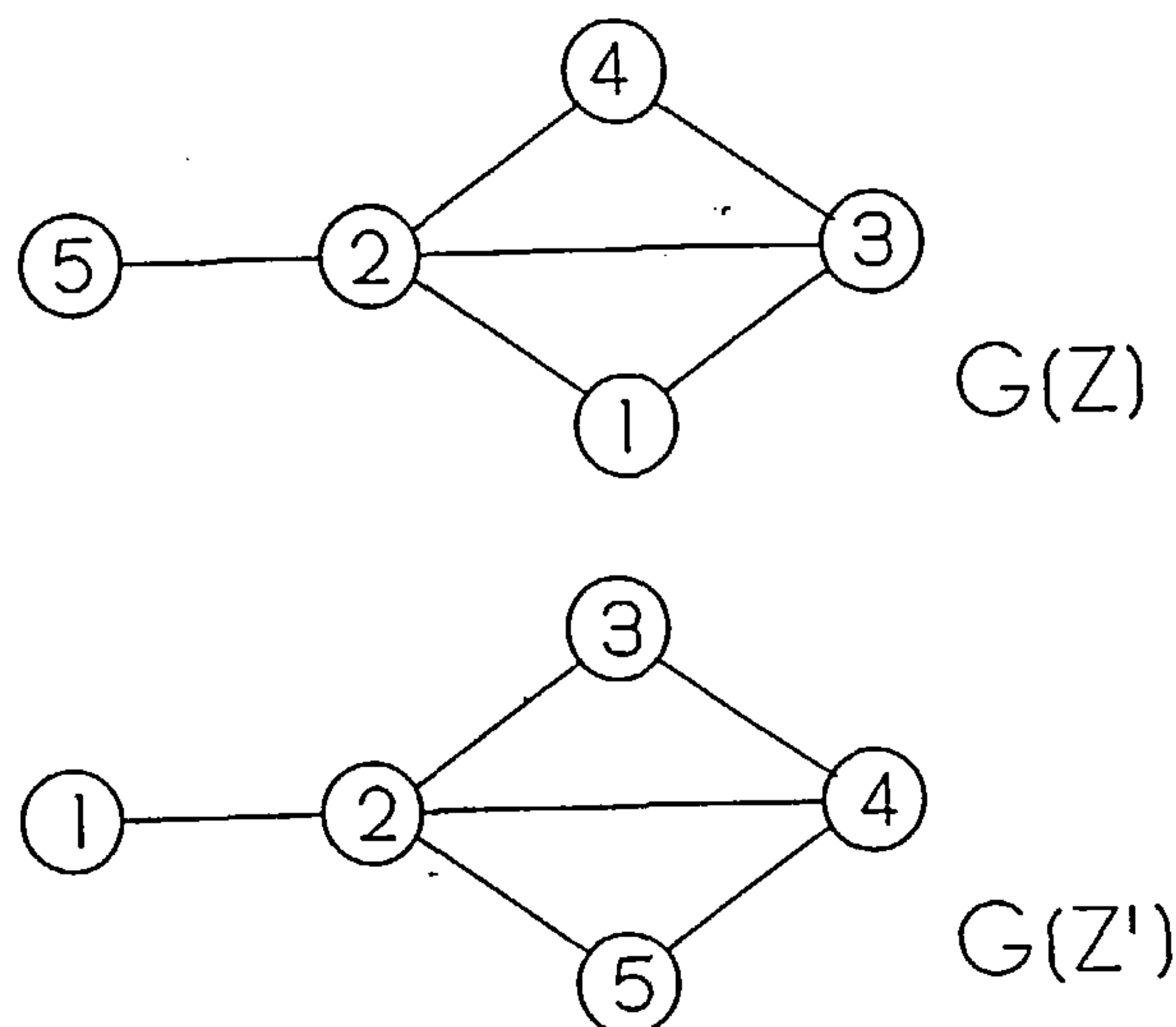


Figure 7.8: $G(Z)$ and $G(Z')$ for the two recursive-directed Z matrices given in equation 7.29.

Notice that only Z' gives a conjugate analysis with two independent Beta distributions on $(\rho'(1), \rho'(2))$.

Thus, although the same graph G can lead to different parameterisations, it is possible to use the graph to find the most appropriate parameterisation to allow a product form Dirichlet prior to posterior analysis. It is shown that the form of the cliques and the order in which they are taken can determine whether a conjugate product prior to posterior Dirichlet exists. The following theorem characterises the conditions required for the reparametrisation of theorem 7.3.2 to take on a product form in terms of $G(Z)$.

Theorem 7.6.1 *The reparametrisation of theorem 7.3.2 has a conjugate product of Dirichlet priors under the d.n.h. iff*

$$\#S(i) = \#C(p(i)) - 1 \quad 2 \leq i \leq n$$

where $\#A$ denotes the number of elements in a set A and $p(i)$ is defined as the

smallest index such that

$$S(i) \left(= C(i) \cap \bigcup_{k=1}^{i-1} C(k) \right) \subseteq C(p(i))$$

where $C(1) \dots C(n)$ are the outcomes sets of Z /cliques of $G(Z)$.

Proof: It is sufficient to show that under (and only under) the conditions above $\dot{\lambda}_i$ is a product of the components $\rho(1) \dots \rho(i-1)$ $i = 2, \dots, n$ where $\dot{\lambda}_i$ and $\rho(k)$ are defined in section 7.3.

To prove that if

$$\#S(i) = \#C(p(i)) - 1 \quad 2 \leq i \leq n$$

then the reparameterisation of λ follows a conjugate of product Dirichlet priors, go by strong induction and assume that the hypothesis is true for $l = 1, \dots, i-1$.

Now suppose that $\#S(i) = \#C(p(i)) - 1$. Then, there is exactly one element j of $C(p(i))$ not in $S(i)$ which has probability

$$\begin{aligned} \dot{\lambda}_{p(i)} \rho_{t_j(p(i))}(p(i)) & \quad \text{if } j \in R(p(i)) = C(p(i)) \setminus S(p(i)) \\ \dot{\lambda}_k \rho_{t_j(k)}(k) & \quad \text{if } j \in S(p(i)), \quad 1 \leq k < p(i) \end{aligned}$$

where $t_j(l)$ denotes the position of element j in $z_l(l)$ and $\rho_{t_j(l)}(l)$ is the corresponding component of $\rho(l)$. Therefore, the equality

$$\dot{\lambda}_i = \dot{\lambda}_k \rho_{t_j(k)}(k) \quad 1 < k \leq p(i)$$

holds. The parameter $\dot{\lambda}_i$ clearly has the required product form provided that the inductive hypothesis holds, so that sufficiency is proved by induction.

It will now be proved that a reparameterisation of a conjugate product of Dirichlet priors must imply that $\#S(i) = \#C(p(i)) - 1$. To prove this, it is necessary to show that whenever i denotes the smallest index such that

$$\#C(p(i)) - \#S(i) \geq 2. \quad (7.30)$$

then $\dot{\lambda}_i$ is not a simple product form of ρ .

Firstly $\dot{\lambda}_i$ is expressed in terms of those λ_j in the different component sets of $S(i)$. It is then shown how, for the various different forms of $C(p(i))$ and $S(i)$ satisfying condition 7.30, $\dot{\lambda}_i$ can never be a simple product form in ρ .

It is clear that:

$$\dot{\lambda}_i = 1 - \sum_{j \in S(i)} \lambda_j. \quad (7.31)$$

$S(i)$ can be partitioned into 2 parts — $S(i) \cap S(p(i))$ and $S(i) \cap R(p(i))$. Equation 7.31 can therefore be rewritten so that:

$$\dot{\lambda}_i = \sigma - \sum_{j \in \{S(i) \cap R(p(i))\}} \lambda_j$$

where

$$\sigma = 1 - \sum_{j \in \{S(i) \cap S(p(i))\}} \lambda_j.$$

The set $R(p(i))$ can be partitioned into the two sets $R(p(i)) \cap S(i)$ and $R(p(i)) \setminus S(i)$.

Thus since

$$\dot{\lambda}_{p(i)}^{-1} \sum_{j \in R(p(i))} \lambda_j = \dot{\lambda}_{p(i)}^{-1} \sum_{j \in \{R(p(i)) \cap S(i)\}} \lambda_j + \dot{\lambda}_{p(i)}^{-1} \sum_{j \in \{R(p(i)) \setminus S(i)\}} \lambda_j = 1,$$

then $\sum_{j \in \{S(i) \cap R(p(i))\}} \lambda_j$ can be rewritten so that:

$$\sum_{j \in \{S(i) \cap R(p(i))\}} \lambda_j = \dot{\lambda}_{p(i)} \tau$$

where

$$\tau = \dot{\lambda}_{p(i)}^{-1} \sum_{j \in \{S(i) \cap R(p(i))\}} \lambda_j = 1 - \sum_{j \in \{R(p(i)) \setminus S(i)\}} \lambda_j / \dot{\lambda}_{p(i)}.$$

Notice that by the definition of $p(i)$, $S(i) \cap R(p(i)) \neq \emptyset$ and so $\tau \neq 0$. $\dot{\lambda}_i$ can then be written in the form:

$$\dot{\lambda}_i = \sigma - \dot{\lambda}_{p(i)} \tau. \quad (7.32)$$

It now remains to show that for all possible cases when

$$\#C(p(i)) - \#S(p(i)) \geq 2$$

$\dot{\lambda}_i$ as expressed in the form in equation 7.32 can never be a product form.

Case I: If $R(p(i)) \setminus S(i)$ contains 2 or more elements and $S(i) \cap S(p(i)) = \emptyset$ then $\sigma = 1$ and τ is the sum of more than 2 items and so clearly $\dot{\lambda}_i$ cannot take the required product form.

Case II: If exactly one element lies in $R(p(i)) \setminus S(i)$, then for $\dot{\lambda}_i$ to take the required product form it is required that:

$$\dot{\lambda}_{(p(i))} = \sigma.$$

However, since $\dot{\lambda}_{p(i)} = 1 - \sum_{j \in S(p(i))} \lambda_j$, this implies that

$$S(i) \cap S(p(i)) = S(p(i)).$$

So

$$\#C(p(i)) - \#S(i) = \#R(p(i)) \setminus S(i) = 1$$

contradicting condition 7.30.

Case III: If $R(p(i)) \setminus S(i) = \emptyset$ so that $R(p(i)) \subseteq S(i)$, then $\tau = 1$ so that equation 7.32 becomes:

$$\dot{\lambda}_i = 1 - \sum_{j \in \{S(i) \cap S(p(i))\}} \lambda_j - \dot{\lambda}_{p(i)}.$$

Since $\dot{\lambda}_{p(i)} = 1 - \sum_{j \in S(p(i))} \lambda_j = \sum_{j \in R(p(i))} \lambda_j$, this becomes:

$$\begin{aligned} \dot{\lambda}_i &= 1 - \sum_{j \in \{S(i) \cap S(p(i))\}} \lambda_j - \sum_{j \in R(p(i))} \lambda_j \\ &= 1 - \sum_{j \in C(p(i))} \lambda_j + \sum_{j \in \{S(p(i)) \setminus S(i)\}} \lambda_j \\ &= \sum_{j \in \{S(p(i)) \setminus S(i)\}} \lambda_j \end{aligned}$$

since $\sum_{j \in C(p(i))} \lambda_j = 1$. Now, it is clear from this equation that for λ_i to have the required product form it is necessary and sufficient that $\#\{S(p(i)) \setminus S(i)\} = 1$. However, since $R(p(i)) \subseteq S(i)$ so that $C(p(i)) = R(p(i)) \cup \{S(p(i)) \setminus S(i)\}$ this breaks condition 7.30 and so gives the required contradiction.

Thus the theorem has been proved.

Suppose there is a recursive-directed matrix Z satisfying the d.n.h. An algorithm is now presented which uses a graph to reparameterise λ so that the reparameterised recursive-directed matrix Z dictates a conjugate Dirichlet prior.

- (1) Form a partition of record nodes into equivalence classes where each node of a given equivalence class is contained in exactly the same set of cliques. (Equivalently collect together the columns of Z which are identical).
- (2) Draw a graph $G^{(1)}$ in the usual way on a set of nodes with exactly one representative from each equivalence class above.
- (3) Label the clique with the largest number of clique intersections $C(1)$. Then introduce an ordering of cliques which satisfy both the RIP and the conditions of theorem 7.6.1. In the case of ties, choose the smallest clique first.
- (4) Form $Z' = Z^{(1)}Z^{(2)} \in G(Z)$ where $Z^{(2)}$ is a partitioning matrix of records representing the equivalence classes above and $Z^{(1)}$ introduces rows in the RIP ordering chosen in (3). The reparametrisation of theorem 7.3.2 is used on the sums of components of λ in each equivalence class and parameters within equivalence classes are introduced as for simple structures.

This procedure gives the simplest parametrisation for simple models and the first example of this section. Here is another example on which this algorithm

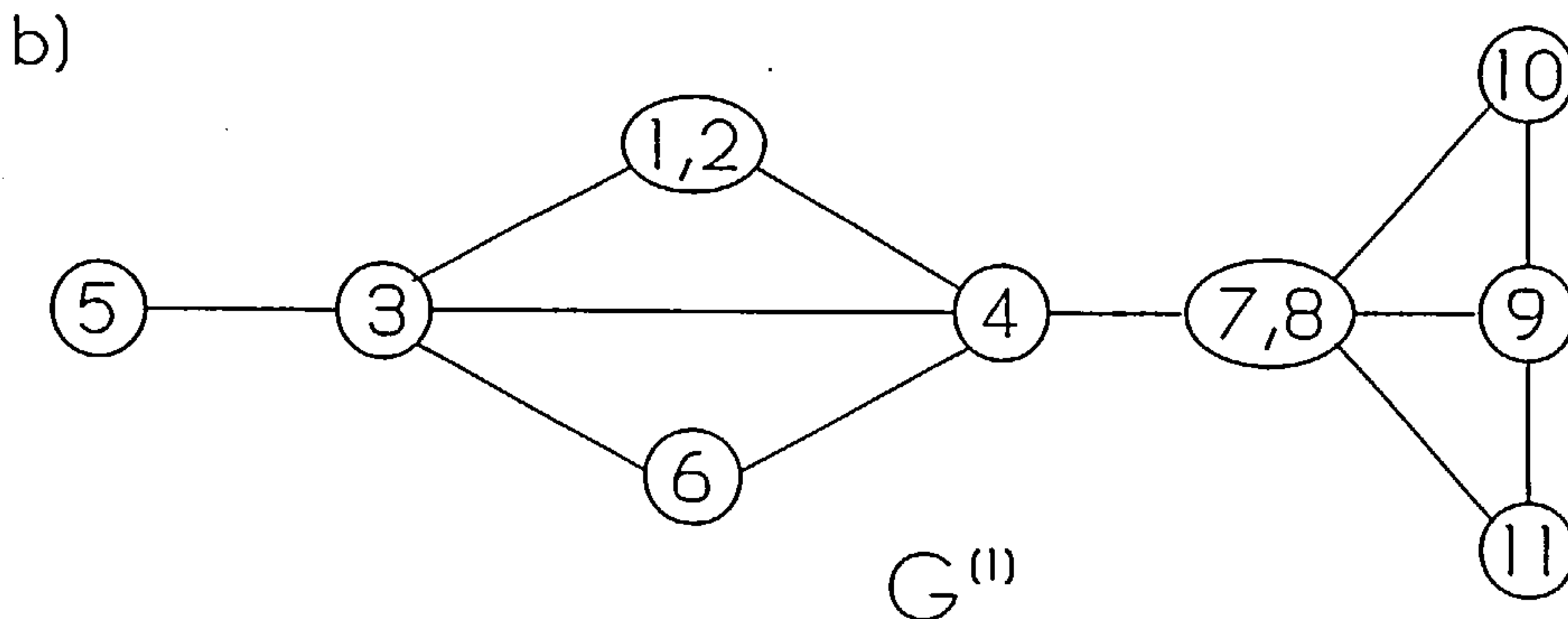
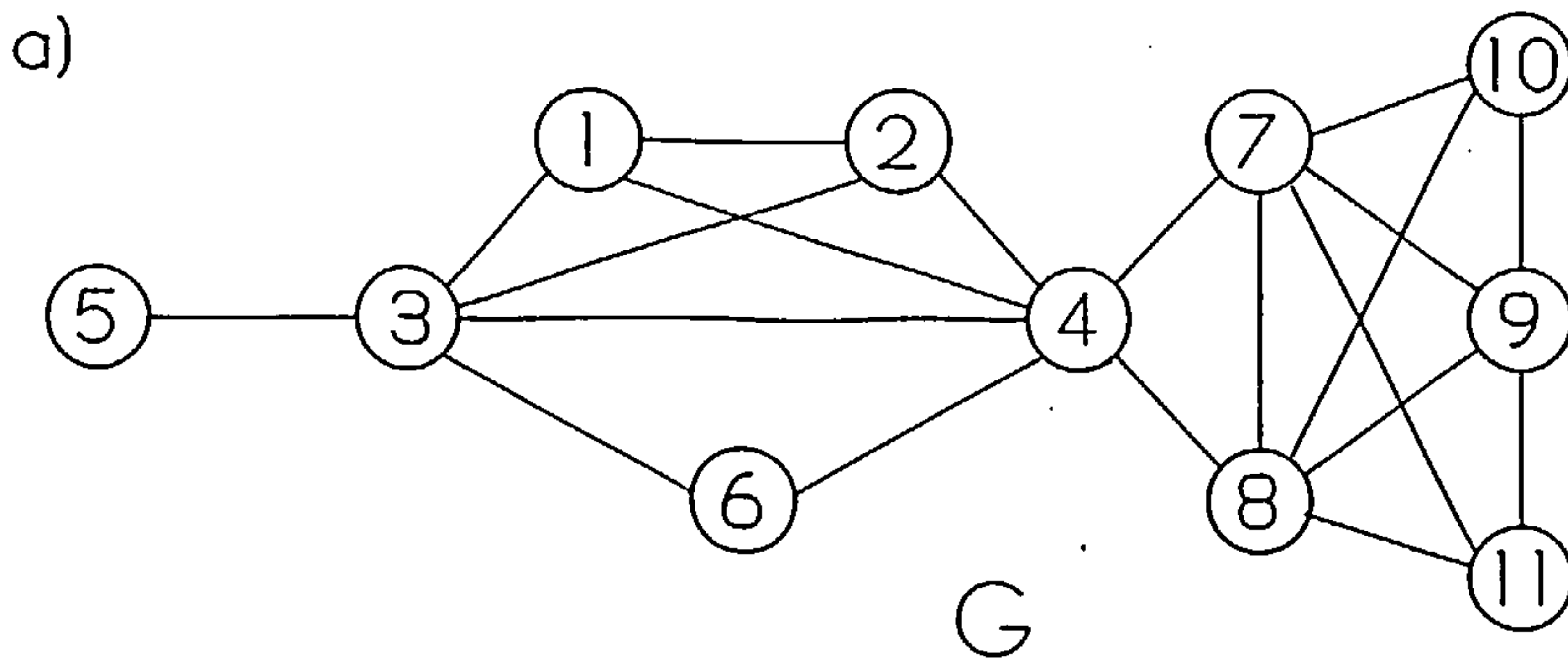


Figure 7.9: Graph G (a) and $G^{(1)}$ (b) such that the nodes on $G^{(1)}$ represent the equivalence classes on G .

works. Let Z be the graphical model given in figure 7.9 (a) whose associated $G^{(1)}$ derived from (2) above is given in figure 7.9 (b). A clique ordering satisfying both the RIP and the conditions of theorem 7.6.1 is given by:

$$C(1) = \{4, (7, 8)\}$$

$$C(2) = \{(7, 8), 9, 10\}$$

$$C(3) = \{(7, 8), 9, 11\}$$

$$C(4) = \{(1, 2), 3, 4\}$$

$$C(5) = \{3, 4, 6\}$$

$$C(6) = \{3, 5\}$$

Borrowing notation from Section 7.3 the reparametrisation of λ then takes

the following form:

$$\begin{aligned}
\lambda_1 &= \rho_2(1)\rho_1(4)\phi_1(1) & \lambda_7 &= \rho_2(1)\phi_1(2) \\
\lambda_2 &= \rho_2(1)\rho_1(4)\phi_2(1) & \lambda_8 &= \rho_2(1)\phi_2(2) \\
\lambda_3 &= \rho_2(1)\rho_2(4) & \lambda_9 &= \rho_2(1)\rho_2(2) \\
\lambda_4 &= \rho_1(1) & \lambda_{10} &= \rho_2(1)\rho_1(2) \\
\lambda_5 &= 1 - \rho_2(1)\rho_2(4) & \lambda_{11} &= \rho_2(1)\rho_1(2) \\
\lambda_6 &= \rho_2(1)\rho_1(4)
\end{aligned}$$

where $\rho_1(i) + \rho_2(i) = 1$, for $i = 1, 2, 4$ and $\phi_1(i) + \phi_2(i) = 1$, for $i = 1, 2$. Notice that although the likelihood on λ does not break down into product form under this reparametrisation. It almost does, the only problem is the contribution $(1 - \rho_2(1)\rho_2(4))^{r_5}$ to the likelihood from the 5th brand. The prior to posterior analysis on $\phi(1)$, $\phi(2)$ and $\rho(2)$ is simply conjugate Beta.

7.7 Conclusion

A well-defined class of censored reporting models with latent local conditional independencies have been identified. It has been shown how these can be reparametrised into a form which allows routine Bayesian prior to posterior analyses, not only of outcome probabilities but also record given outcome probabilities. This simplification is very useful—for example it allows time series models to be built which respect the conditional independencies defining a problem.

The structures which allow simple reparametrisations are closely linked with classes of decomposable graphs. These graphs not only give a concise description of a model but can also guide in the discovery of appropriate reparametrisations.

Throughout this chapter it has been assumed that the matrix Z of likelihood ratios is known. However powerful tests of a hypothesised Z against certain classes of alternative exist. It is also possible to estimate some of the parameters

of Z , given that Z is recursive-directed. These topics will be addressed in a later paper.

Chapter 8

Discussion and further research.

This thesis has developed three new classes of Bayesian multivariate forecasting model — namely the Multiregression Dynamic Model (chapter 5), the Dynamic Graphical Model (chapter 6) and the Partial Segmentation Model (chapter 7). Although all the models were initially motivated by the need to develop comprehensive models of competitive business markets, MDM's and DGM's considered the problem from a different angle to that concentrated on by Partial Segmentation Models. The MDM and DGM modelled any conditional causal structures imposed by both the partial segmentation and aggressive competitive strategies within markets, whereas the Partial Segmentation Model concentrated on modelling the partial segmentation directly.

MDM's/DGM's do not rely on the stringent symmetry conditions imposed by the multivariate Bayesian forecasting models of Harvey (1986) or the DMR model (see section 3.6) and the Partial Segmentation Model provides an alternative to the Dirichlet model (see section 2.4). However, the classes of MDM and DMG are still not rich enough to capture *all* the different types of dependence between brand sales which might be expected. Further research is therefore required to extend these models to accommodate more dependence structures. Further research is also required to address the issue of model discrimination so that diagnostic

checks for these models can be formulated before the models can be implemented in practice. It therefore seems hopeful that these models can provide a foundation for further research into the practical problem of modelling competitive business markets.

The models are also of theoretical interest. MDM's/DGM's offer a relatively simple method of modelling highly non-Gaussian multivariate time series and also provide a vehicle by which heuristic causal links and conditional independence structures between components in a multivariate series can be accommodated. Partial Segmentation Models extend the work of Dickey et al. (1987) so that censored categorical data with partial segmentation (as opposed to the simple segmentation which Dickey et al. consider) can be modelled. They also link these categorical models to graphs which, not only provide a good pictorial representation of the market structure, but can also be used as a guide by which the most appropriate parameterisation for the model of a particular market can be found.

As was mentioned earlier, the MDM/DGM and the Partial Segmentation Model were developed by considering competitive business markets from slightly different angles. The next section shows how the Partial Segmentation Model for a particular market can be generalised into a form compatible with the MDM and DGM structures.

8.1 Integration of Partial Segmentation Models and MDM's/DGM's

Two examples will be presented to illustrate how the Partial Segmentation Model and the MDM/DGM can be integrated.

Suppose that a market has just 4 brands and two types of customer. Suppose further that the differentially non-informative hypothesis holds so that the

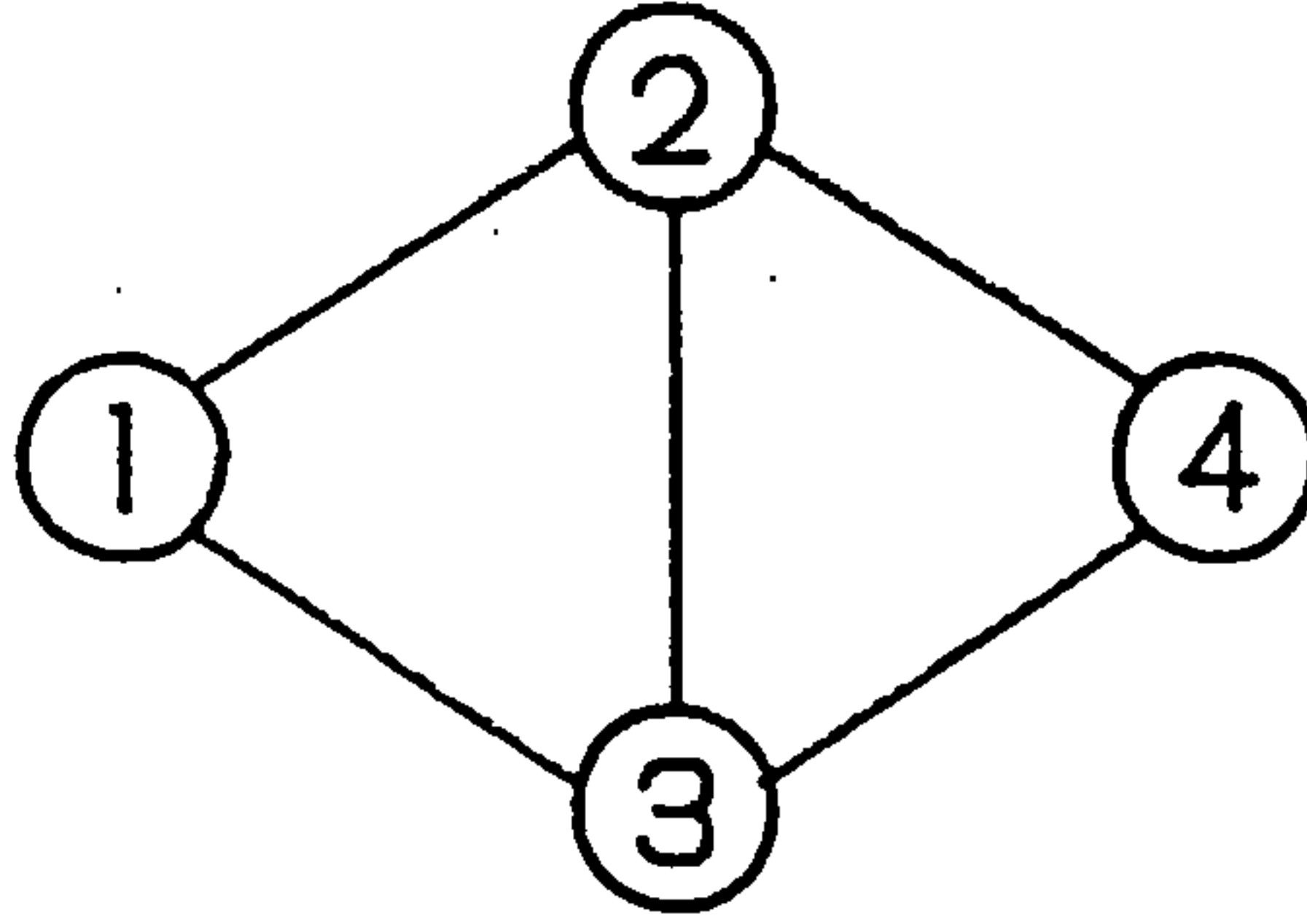


Figure 8.1: $G(Z)$ for Z in equation 8.1.

hypothesised likelihood ratio recursive-directed matrix Z for this market is given by:

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix}. \quad (8.1)$$

The graph $G(Z)$ can be seen in figure 8.1.

Using the notation of section 7.3 notice that:

$$\psi_1 = \theta(1)\rho_1(1)$$

$$\psi_2 = \rho_2(1)\rho_1(2)$$

$$\psi_3 = \rho_2(1)\rho_2(2)$$

$$\psi_4 = \theta(2)\rho_1(1).$$

where $\theta(1) + \theta(2) = 1$ and $\rho_1(i) + \rho_2(i) = 1$ for $i = 1, 2$.

Now make the assumption that observed sales of the 4 brands (R_1, R_2, R_3, R_4) have the following independent distributions *given* their parameters:

$$R_1 \sim Po(\mu\theta(1)\rho_1(1))$$

$$R_2 \sim Po(\mu\rho_2(1)\rho_1(2))$$

$$R_3 \sim Po(\mu\rho_2(1)\rho_2(2))$$

$$R_4 \sim Po(\mu\theta(2)\rho_1(1)).$$

where $Po(\alpha)$ denotes the Poisson distribution with mean α . The likelihood for the

parameters μ , θ and ρ for this model is simply the product of the 4 independent likelihoods and is given by:

$$\begin{aligned} L(\mu, \theta, \rho | \mathbf{r}) &= \exp [-\mu \{ \theta(1)\rho_1(1) + \rho_2(1)\rho_1(2) + \rho_2(1)\rho_2(2) + \theta(2)\rho_1(1) \}] \\ &\quad \times \{ \mu\theta(1)\rho_1(1) \}^{r_1} \{ \mu\rho_2(1)\rho_1(2) \}^{r_2} \{ \mu\rho_2(1)\rho_2(2) \}^{r_3} \{ \mu\theta(2)\rho_1(1) \}^{r_4} \\ &= L_1(\theta)L_2(\rho)L_3(\mu) \end{aligned}$$

where

$$\begin{aligned} L_1(\theta) &= \theta(1)^{r_1}\theta(2)^{r_4} \\ L_2(\rho) &= \rho_1(1)^{r_1+r_4}\rho_2(1)^{r_2+r_3}\rho_1(2)^{r_2}\rho_2(2)^{r_3} \\ L_3(\mu) &= \mu^N e^{-\mu} \end{aligned}$$

such that $N = \sum_{j=1}^4 r_j$. Note that for a fixed value of N this is the likelihood which would be obtained assuming that \mathbf{R} followed a multinomial distribution and θ , $\rho(1)$, $\rho(2)$ had the generalised Dirichlet structure described in chapter 7.

For any two variables X_1 and X_2 such that $X_1 \sim Po(\mu_1)$ and $X_2 \sim Po(\mu_2)$, the distribution of $\{X_1 | X_1 + X_2\}$ is binomial so that:

$$(X_1 | X_1 + X_2) \sim Bi \left(X_1 + X_2, \frac{\mu_1}{\mu_1 + \mu_2} \right)$$

where $Bi(n, p)$ denotes the binomial distribution from a sample of n with parameter p . Using this result the independent brand sales (R_1, \dots, R_4) can be transformed to give the following 4 variables:

$$\begin{aligned} N &\sim Po(\mu) \\ (R_1 + R_4 | N) &\sim Bi(N, \rho_1(1)) \\ (R_2 | N - (R_1 + R_4)) &\sim Bi(N - (R_1 + R_4), \rho_1(2)) \\ (R_1 | R_1 + R_4) &\sim Bi(R_1 + R_4, \theta(1)). \end{aligned}$$

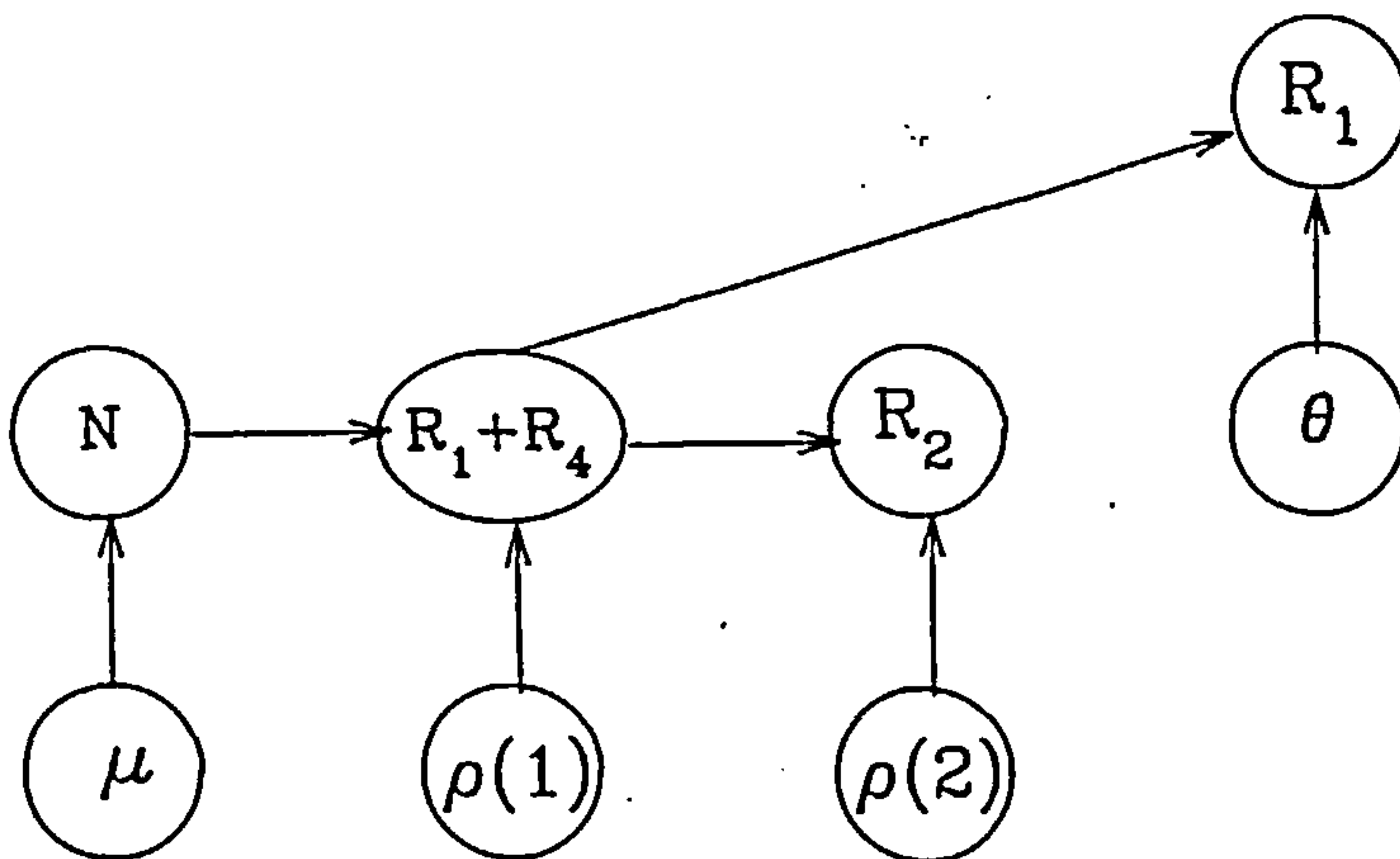


Figure 8.2: Graph of the ID representing the transformed series of brand sales R .

The independent series R have therefore been transformed to a set of series for which a conditional independence structure can be defined across. This conditional independence structure is represented by the graph of the influence diagram of figure 8.2. This graph defines the structure of an MDM for the series

$$\{Y(1), Y(2), Y(3), Y(4)\} = \{N, R_1 + R_4, R_2, R_1\}$$

and

$$\{\theta(1), \theta(2), \theta(3), \theta(4)\} = \{\mu, \rho(1), \rho(2), \theta\}$$

such that μ , $\rho(1)$, $\rho(2)$ and θ are a priori independent.

Notice that these are not normal MDM's but the conditional independence results still hold. The obvious choice for models of this type would be to make the conditional components of the MDM have a DGLM (see West & Harrison, 1989a) structure which would allow for the regression seasonality and trend variables to be modelled in a larger state space. Of course the predictive distribution would be a rather complicated product of the Poisson-Gamma/Beta-Binomial conditional distributions on the margins of Y . However, the predictive moments would be

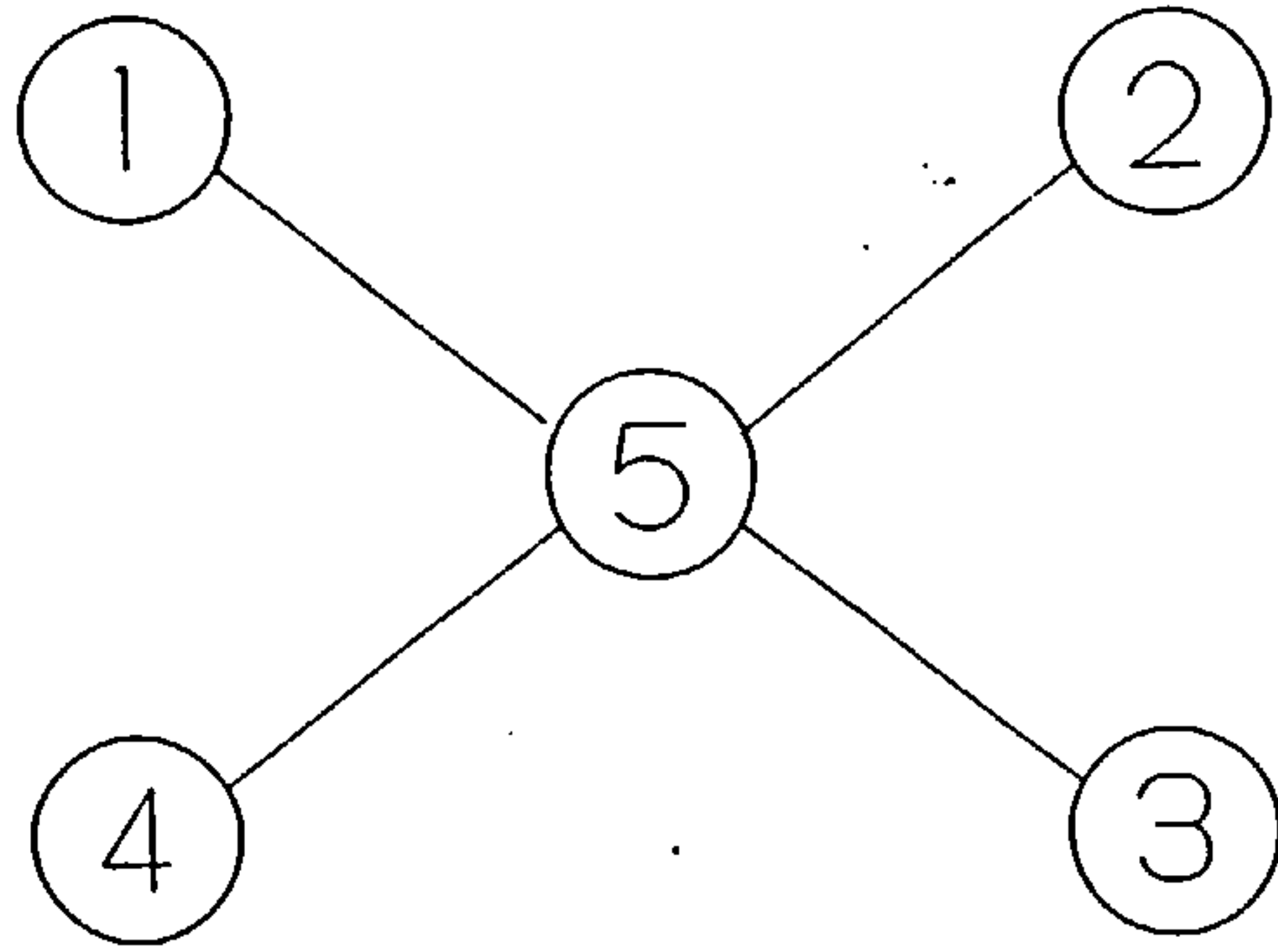


Figure 8.3: Graph $G(Z)$ representing the hypothesised likelihood ratios in a star model for 5 brands.

fairly simple to calculate as for the LMDM. Time however prohibits the study and application of such processes which will be studied in a later paper.

A commonly assumed partial segmentation used by Unilever, is what shall be called a *star model* here. In this case the market is assumed to be structured so that every purchaser will choose between a “favourite brand” or one other alternative.

Consider a hypothetical example of such a market in which there are 5 brands and brand 5 is the favoured brand. Suppose that the graph $G(Z)$ for the hypothesised likelihood ratio matrix Z satisfying the d.n.h. is given in figure 8.3.

Once again using the notation of section 7.3 notice that:

$$\begin{aligned}\psi_j &= \theta(j)\lambda_1, & \text{for } j = 1, \dots, 4, \\ \psi_5 &= 1 - \lambda_1.\end{aligned}$$

where $\sum_{j=1}^4 \theta(j) = 1$. In a similar fashion to the first example, assume that the sales of the 5 brands (R_1, \dots, R_5) have the following independent distributions given their parameters:

$$R_1 \sim Po(\mu\lambda_1\theta(1))$$

$$\begin{aligned}
R_2 &\sim Po(\mu\lambda_1\theta(2)) \\
R_3 &\sim Po(\mu\lambda_1\theta(3)) \\
R_4 &\sim Po(\mu\lambda_1\theta(4)) \\
R_5 &\sim Po(\mu(1 - \lambda_1)).
\end{aligned}$$

The likelihood for the parameters μ , λ_1 and θ is then given by:

$$\begin{aligned}
L(\mu, \lambda_1, \theta | r) &= \exp[-\mu\{\lambda_1(\theta(1) + \theta(2) + \theta(3) + \theta(4)) + (1 - \lambda_1)\}] \\
&\quad \times \{\mu\theta(1)\lambda_1\}^{r_1} \{\mu\theta(2)\lambda_1\}^{r_2} \{\mu\theta(3)\lambda_1\}^{r_3} \{\mu\theta(4)\lambda_1\}^{r_4} \{\mu(1 - \lambda_1)\}^{r_5} \\
&= L_1(\theta)L_2(\lambda_1)L_3(\mu)
\end{aligned}$$

where

$$\begin{aligned}
L_1(\theta) &= \theta(1)^{r_1} \theta(2)^{r_2} \theta(3)^{r_3} \theta(4)^{r_4} \\
L_2(\lambda_1) &= \lambda_1^{\sum_{j=1}^4 r_j} (1 - \lambda_1)^{r_5} \\
L_3(\mu) &= \mu^N e^{-\mu}
\end{aligned}$$

such that $N = \sum_{j=1}^5 r_j$.

As for the previous example, transform the independent series R to a set of series which have a conditional independence structure defined across them. The transformed variables then have the following distributions:

$$\begin{aligned}
N &\sim Po(\mu) \\
N - R_5 &\sim Bi(N, \lambda_1) \\
(R_1, R_2, R_3, R_4) &\sim Mn\{N - R_5, \{\theta(1), \theta(2), \theta(3), \theta(4)\}\}
\end{aligned}$$

where $Mn(n, p)$ denotes the multinomial distribution for sample size n with parameter vector p . Assuming μ , θ , λ_1 are apriori independent, then the conditional

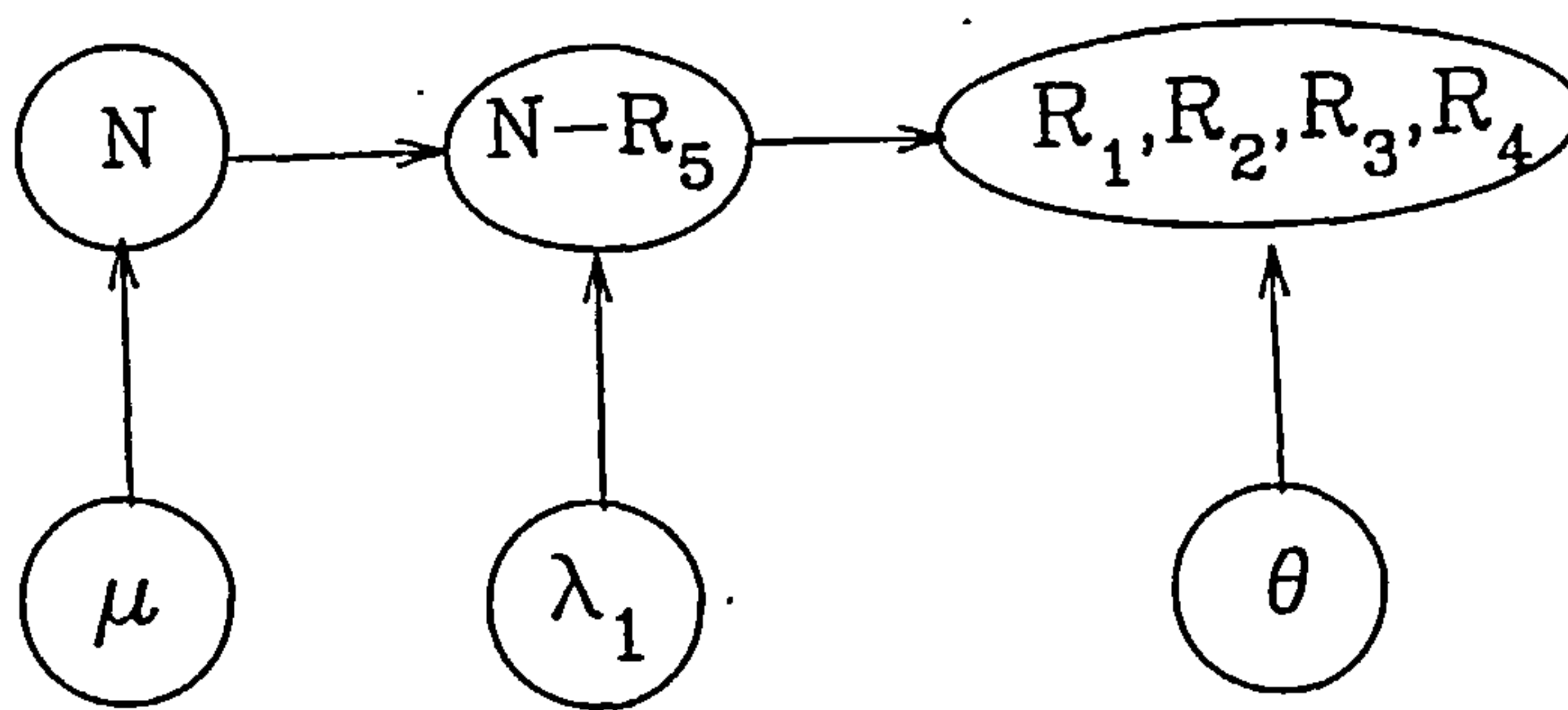


Figure 8.4: Graph of the ID representing the transformed variables of 5 brands with a star model.

independencies which exist between these transformed variables can be expressed by the graph of the influence diagram of figure 8.4.

Once again the transformed variables represent the conditional independence structure of an MDM, so that the iterative structure still applies. The interest in this process lies in the structure of (R_1, R_2, R_3, R_4) — a vector regressed on a sum which is similar to the form of a DGM. Again multivariate DGLM's (Atwell & Smith, 1990) can be used to construct a full model for this process.

In the applications which originally motivated these studies, the value of total sales, represented by N in the last two examples, would be enormous. In these cases, the posterior variance of each parameter after having observed R is negligible and so, without loss, each parameter can be identified with its maximum likelihood estimate. Thus for the first example, the dependent variables

$$\begin{aligned} (Y(1), \dots, Y(4)) &= \left(N, \frac{R_1 + R_4}{N}, \frac{R_2}{N - (R_1 + R_4)}, \frac{R_1}{R_1 + R_4} \right) \\ &\approx (\mu, \rho_1(1), \rho_1(2), \theta(1)). \end{aligned}$$

However, it was assumed that the parameters μ , $\rho(1)$, $\rho(2)$ and θ are a priori independent. Thus with a partial segmentation and the d.n.h. hypothesis it is expected that when N is very large, the processes $Y(1), \dots, Y(4)$ are independent

and so the dependence degenerates and so an MDM structure is no longer apparent. It appears that $Y(1), \dots, Y(4)$ can be modelled separately with univariate models and the joint distributions of (R_1, \dots, R_4) be obtained from the inverse transform of the one above.

However, there are two main reasons why the MDM structure will be important here. Firstly, when one or more players are being aggressive about their marketing policy, the d.n.h. is unlikely to hold in the long term, since it is unlikely that all types who buy a certain brand have an equal probability of buying that brand over time. Secondly, it was shown in section 5.5 when demonstrating the differences between the interpretations of the LMDM and the CLMDM that a plausible CLMDM for an ice-cream market would regress on the aggregate of sales to allow for differences in adaptability between the different brand sales.

Again time constraints have not allowed for the systematic study of such processes which estimate parameters of segmentation on-line and relate these to the actions of the various competitors. This will be discussed at length in a forthcoming paper. Notice that in the example of the star model, for large N , (R_1, R_2, R_3, R_4) would be conveniently modelled using the logistic transform and the DMR model (see section 3.6) linking their models to the DGM.

8.2 Other topics for Further Research.

Sales data is often insufficient to identify quantities of interest about the dynamics of a market. However, surveys, like the panel data presented in section 2.1, can be investigated and these can often provide useful information. Suitable models of the contingency table of panel data which exhibit the appropriate forms of conditional independence structure is sympathetic to an analysis by Hyper Markov Dirichlet Processes (Dawid & Lauritzen, 1990). It has already

been established how such Dirichlet models can be simply stochasticised (see, for example, Smith, 1981, Attwell & Smith, 1989). It is necessary to study how these models perform in this application and also to incorporate the information from such a study into the formulation of MDM's, DGM's and Partial Segmentation Models.

Before it is possible to produce models which realistically take into account the effects of competitive strategies on the various brand sales in the market, it is necessary to predict how companies are likely to react to one another. An obvious starting point in such an analysis is to assume that individual companies will react rationally (see Smith & Young, 1988, Young & Smith, 1991a, 91b). It can be shown (see, for example, Smith, 1988) that these models whose conditional independence structures are consistent with mutual rationality, can be identified at least in simple stochastic games. It is hoped that these ideas can be extended to the new classes of models to identify interesting subclasses. It should be possible, through linking Granger causality with conditional independence structures, to analyse and test the ramifications of the postulated rational control by players on the causal structure.

Appendix A

Consistency and Estimability of Partially Segmented Markets.

To consider whether the posterior distribution of θ converges as $N \rightarrow \infty$ the likelihood function needs to be inspected. Let $N = \sum_{j=1}^m r_j$ and $x = (x_1, \dots, x_m)^T$ where x_j is defined by $Nx_j = r_j$, $1 \leq j \leq m$. Then

$$\begin{aligned} L_1(\theta) &= \prod_{j=1}^m (\xi_j(\theta))^{r_j} \\ &= \exp \left(\log \left(\prod_{j=1}^m \xi_j(\theta)^{r_j} \right) \right) \\ &= \exp \left(\sum_{j=1}^m \log (\xi_j(\theta)^{r_j}) \right) \\ &= \exp \left(\sum_{j=1}^m r_j \log (\xi_j(\theta)) \right) \end{aligned}$$

but $r_j = Nx_j$ and so

$$L_1(\theta) = \exp \{N.l_1(\xi(\theta))\} \quad \xi(\theta) = (\xi_1(\theta), \dots, \xi_m(\theta)) \in A \quad (\text{A.1})$$

where

$$l_1(\xi(\theta)) = \sum_{j=1}^m x_j \log \xi_j(\theta), \quad 0 < \xi_j(\theta) < 1, \quad 1 \leq j \leq m.$$

The set A given in equation A.1 constraining the vector ξ is a convex subspace of \Re^{n-1} . To prove this, consider partitioning $\xi(\theta)$ and Z . Let $\xi(\theta)^T = [\xi_1(\theta)^T, \xi_2(\theta)^T]$, where $\xi_1(\theta)$ is the first n elements of $\xi(\theta)$ and $\xi_2(\theta)$ is the remaining $m - n$. Let $Z = [Z_1, Z_2]$ where Z_1 is the first n columns of Z and Z_2 is the remaining $m - n$ columns. Since Z is of rank n , without loss of generality, the components of ξ can be reordered so that Z_1 is non-singular. Thus equation 7.7 can be rewritten as:

$$\begin{bmatrix} \xi_1(\theta) \\ \xi_2(\theta) \end{bmatrix} = \begin{bmatrix} Z_1^T \\ Z_2^T \end{bmatrix} \theta$$

and so

$$\xi_1(\theta) = Z_1^T \theta.$$

Now, the constraint $\sum_{i=1}^n \theta(i) = 1$ means that the solution space for $\xi_1(\theta)$ (and hence $\xi(\theta)$) is contained in \Re^{n-1} . The set A is then simply the intersection of the n convex subspaces of \Re^{n-1} imposed by the further constraints $\theta(i) > 0$, $1 \leq i \leq n$.

It is now necessary to show that $l_1(\xi(\theta))$ is a strictly concave function. Now, as A is a convex set, $l_1(\xi(\theta))$ will also be strictly concave when restricted to A and so there is a unique value $\hat{\xi}(\theta) \in \bar{A}$, the closure of A , which maximises $l_1(\xi(\theta))$ on \bar{A} . This ensures there is only one local maximum in the likelihood. In particular

$$\frac{\partial^2 l_1(\xi(\theta))}{\partial \xi_j^2(\theta)} = \frac{-x_j}{\xi_j^2(\theta)} < -x_j \leq -\dot{x} < 0 \quad \text{for } 0 < \xi_j < 1, 1 \leq j \leq m \quad (\text{A.2})$$

where $\dot{x} = \min_{1 \leq j \leq m} \{x_j\}$. If for some $\epsilon > 0$, $\mathbf{x} \in B(\epsilon)$ where

$$B(\epsilon) = \{\mathbf{x} : \dot{x} \geq \epsilon\},$$

then, as the second derivative of $l_1(\xi(\theta)) < 0$, it follows that $l_1(\xi(\theta))$ is a strictly concave function.

To prove that the posterior distribution of θ degenerates to $\hat{\xi}(\theta)$ it is sufficient to show that under a sufficiently non-degenerate prior, $L_1(\theta)$ degenerates to $\hat{\xi}(\theta)$ as $N \rightarrow \infty$. Clearly from equations A.1 and A.2, for all $\xi(\theta) \in \bar{A}$, $\|\hat{\xi}(\theta) - \xi(\theta)\| > \eta$, $\eta > 0$, provided that $x \in B(\epsilon)$ for some $\epsilon > 0$,

$$l_1(\hat{\xi}(\theta)) - l_1(\xi(\theta)) > 0$$

and therefore

$$N [l_1(\hat{\xi}(\theta)) - l_1(\xi(\theta))] \rightarrow \infty, \quad \text{as } N \rightarrow \infty$$

This ensures that provided $x \in B(\epsilon)$ for some $\epsilon > 0$ under any bounded prior density $p(\xi(\theta))$ on $\xi(\theta) \in \bar{A}$, non-zero anywhere on \bar{A} , for all $\eta > 0$

$$P(\|\hat{\xi}(\theta) - \xi(\theta)\| \geq \eta) \rightarrow 0 \quad N \rightarrow \infty$$

where P is the probability distribution associated with a maximum likelihood estimate $\hat{\xi}(\theta)$ averaged over the prior distribution on $\xi(\theta)$.

Thus if $x \in B(\epsilon)$ for some $\epsilon > 0$, the posterior distribution of $\xi(\theta)$ will converge in distribution to unit mass of $\hat{\xi}(\theta)$ — i.e., $\xi(\theta)$ will become consistently Bayes estimable. Assume that the random variable of records r is, by hypothesis, multinomially distributed $M(N, \psi)$ where N is defined above and $\psi = (\psi_1, \dots, \psi_m)^T$

$$\psi_j = \lambda_j \xi_j(\theta) > 0, \quad 1 \leq j \leq m.$$

It follows that if $\dot{\psi} = \min_{1 \leq j \leq m} \psi_j$ and $\dot{\psi} \geq \delta$ then, by the law of large numbers, for all $\epsilon < \epsilon(\delta)$,

$$P\{X(N) \in B(\epsilon) \mid \psi : \dot{\psi} \geq \delta\} \rightarrow 1 \quad \text{as } N \rightarrow \infty$$

where $X(N)$ is the random variable representing the proportions of records. So, for any fixed (λ, θ) ($\lambda_j > 0$, $\theta_j > 0$, $1 \leq j \leq m$), for any $\epsilon > 0$, if N can be made sufficiently large then $X(N) \in B(\epsilon)$ is virtually certain.

It follows that, under a sufficiently non-degenerate prior distribution, the posterior distribution of θ will converge to a point $\hat{\theta}$ if Z is of rank n . On the other hand, if Z is not of rank n , since $L_1(\theta)$ is only a function of θ through $\xi(\theta)$, linear ridges on $L_1(\theta)$ will exist and the posterior distribution of θ , based on r , will not converge.

References.

- Ameen, J.R.M. (1984) *Discount Bayesian Models and Forecasting*. Unpublished PhD thesis, University of Warwick.
- Ameen, J.R.M. & Harrison, P.J. (1985a) Normal discount Bayesian models. In *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.). North-Holland, Amsterdam and Valencia University Press.
- Ameen J.R.M. & Harrison, P.J. (1985b) Discount Bayesian multi-process modelling with CUSUMS. In *Time Series Analysis: Theory and Practice*, 5, O.D. Anderson (Ed.). North-Holland, Amsterdam.
- Attwell, D.N. & Smith, J.Q. (1989) A Bayesian forecasting model for sequential bidding. *Research Report 170*, Department of Statistics, University of Warwick.
- Attwell, D.N. & Smith, J.Q. (1990) A dynamic generalised linear model for sequential bidding. *Research Report 207*, Department of Statistics, University of Warwick.
- Beeri, C., Fagin, R., Majer, D., Mandelzon, A., Ullman, J. & Yannakakis, M. (1981) Properties of acyclic database shemes. In *Proc. 13th Annual ACM Symposium on the Theory of Computing, Milwaukee*. New York: Association of Computing Machines.
- Beeri, C., Fagin, R., Majer, D. & Yannakakis, M. (1983) On the desirability of acyclic database schemes. *J. Ass. Comput. Mach.*, 30, 479-513.
- Berge, C (1973) *Graphs and hypergraphs*. Transl. from French by E. Minieka. Amsterdam. North- Holland.
- Bourgeois, J.C., Haines, G.H. & Sommers, M.S. (1980) Defining an industry. In *Market Measurement and Analysis*, D.B. Montgomery & G. Wittink (Eds.).

Cambridge, MA: Marketing Science Institute.

Bourgeois, J.C., Haines, G.H. & Sommers, M.S. (1982) Product/market structures: problems and issues. In *Analytic Approaches to Product and Market Planning*, R.K. Srivastava & A.D. Shocker (Eds.). Cambridge, MA: Marketing Science Institute.

Box, G.E.P. & Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*, (2nd edn.). Holden-Day, San Fransisco.

Brown, R.G. (1959) *Statistical Forecasting for Inventory Control*. McGraw-Hill, New York.

Brown, R.G. (1962) *Smoothing, Forecasting and Prediction of Discrete Time Series*. Prentice-Hall, Englewood Cliffs, NJ.

Carlson, B.C. (1977) *Special Function of Applied Mathematics*. New York: Academic Press.

Chan, W.Y.T. & Wallis, K.F. (1978) Multiple time series modelling: another look at the mink-muskrat interaction. *Applied Statistics*, **27**, 168–175.

Chatfield, C & Collins, A.J. (1990) *Introduction to Multivariate Analysis*. Chapman & Hall Ltd.

Colman, S. & Brown, G. (1983) Advertising tracking studies and sales effects. *J. Market Res. Soc.*, **25**, 165–183.

Dawid, A.P. (1979) Conditional independence in statistical theory. *J.R. Statist. Soc. B*, **41**, 1–31.

Dawid, A.P. (1981) Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**, 265–274.

Dawid, A.P. & Dickey, J.M. (1977) Likelihood and Bayesian inference from selectively reported data. *J. of Amer. Statist. Ass.*, **72**, 845–850.

Dawid, A.P. & Lauritzen, S.L. (1990). Markov distributions, hyper Markov

laws and meta Markov models on decomposable graphs, with applications to Bayesian learning in expert systems. *Research Report R89-31*, University of Aalborg.

Day, G.S., Shocker, A.D. & Srivastava, R. (1979). Customer-oriented approaches to identifying product-markets. *J. of Marketing*, 43, 8-19.

Dechter, R. & Pearl, J. (1987) Network-based heuristics for constraint-satisfaction problems. *Artificial Intelligence*, 34, 1-38.

Dechter, R., Dechter, A. & Pearl, J. (1990) Optimisation in constraint networks. In *Influence Diagrams, Belief Nets and Decision Analysis*. R.M. Oliver & J.Q. Smith (Eds.) John Wiley & Sons Ltd., 411-425.

Dickey, J.M. (1983) Multiple hypergeometric functions: probabilistic interpretations and statistical uses. *J. of Amer. Statist. Ass.*, 78, 628-637.

Dickey, J.M., Jiang, J. & Kadane, J.B. (1987) Bayesian methods for censored categorical data. *J. Amer. Statist. Ass.*, 82, 399, 773-781.

Dirac, G.A. (1961) On rigid circuit graphs. *Abh. Math. Sem. Univ. Hamburg*, 25, 71-76.

Engle, R.F. & Granger, C.W.J. (1987) Co-integration and error correction: representation, estimation and testing. *Econometrica*, 55, 251-276.

Florens, J.P. and Mouchart, M. (1985) A linear theory for non-causality. *Econometrica*, 53, 157-175.

Frank, R.E., Massy, W.F. & Wind, Y. (1972) *Market Segmentation*. Englewood Cliffs (N.J.), Prentice-Hall.

Frydenberg, M (1989) The chain graph Markov property. *Research Report 186*, Department of Theoretical Statistics, University of Aarhus.

Gavril, T. (1972) Algorithms for minimum colouring, maximum clique, minimum colouring by cliques and maximum independent set of a chordal graph.

SIAM J. Comput., **1**, 180–187.

Golumbic, M.C. (1980) *Algorithmic Graph theory and Perfect Graphs*. London: Academic Press.

Goodhardt, G.J., Ehrenberg, A.S.C. & Chatfield, C. (1984) The Dirichlet: a comprehensive model of buyer behaviour. *J.R. Statist. Soc. A*, **147**, Part 5, 621–655.

Granger, C.W.J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Granger, C.W.J. (1980) Testing for causality : a personal viewpoint. *Journal of Economic Dynamics and Control*, **2**, 329–352.

Hannan, E.J. & Kavalieris, L. (1984) Multivariate linear time series models. *Adv. Appl. Prob.*, **16**, 492–561.

Harrison, P.J. and Stevens, C.F. (1976) Bayesian forecasting (with discussion). *J.R. Statist. Soc. B*, **38**, 205–247.

Harvey, A.C. (1986) Analysis and generalisation of a multivariate exponential smoothing model. *Management Science*, **32** (3), 374–380.

Harvey, A.C. (1989) *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Harvey, A.C. & Stock, J.H. (1988) Continuous time autoregressive models with common stochastic trends. *J. of Economic Dynamics and Control*, **12**, 365–384.

Holland, P.W. (1986) Statistics and causal inference. *J. of the Amer. Statist. Soc.*, **81** (396), 945–970.

Holt, C.C. (1957) Forecasting seasonals and trends by exponentially weighted averages. *O.N.R. Research Memo*, **52**, Carnegie Institute of Technology.

Howard, R.A. and Matheson, J.E. (1981) Influence Diagrams. In *Readings*

on the principles and applications of decision analysis, Vol II, R.A. Howard and J.E. Matheson (Eds.) Strategic Decision Group, Menlo Park, Calif., 719–762.

Jewell, N.P. & Bloomfield, P. (1983) Canonical correlations of past and future for time series: definitions and theory. *Annals of Statistics*, **11**, 837–847.

Jewell, N.P., Bloomfield, P. & Bartmann, F.C. (1983) Canonical correlations of past and future for time series: bounds and computation. *Annals of Statistics*, **11**, 848–855.

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. *J. of Basic Engineering*, **82**, 35–45.

Kalman, R.E. (1963) New methods in Wiener filtering theory. In *Proceedings of the First Symposium of Engineering Applications and Random Function Theory and Probability*. J.L. Bogdanoff & F. Kozin (Eds.) Wiley, New York.

Kiiveri, H., Speed, T.P. & Carlin, J.B. (1984) Recursive causal models. *J. Austral. Math. Soc., A*, **36**, 30–51.

Lauritzen, S.L. (1989) Mixed graphical associated models. *Scand. J. Statist.*, **16**, 273–306.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N. and Leimer, H.G. (1990) Independence properties of directed markov fields. *Networks*, **20**, 491–505.

Lauritzen, S.L., Speed, T.P. & Vijayan, K. (1984) Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. A.*, **36**, 12–29.

Lauritzen, S.L. & Spiegelhalter, D.J. (1988) Local computations with probabilities on graphical structures and their applications to expert systems. *J. Royal Statist. Soc., B*, **50**, 2, 157–224.

Lauritzen, S.L. & Wermuth, N. (1984) Mixed interaction models. *Research Report R-84-8*. Inst. of Elec. Sys., University of Aalborg.

Lauritzen, S.L. & Wermuth, N. (1989) Graphical models for associations be-

tween variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17, 31–57.

Maybeck, P.S. (1979, 1982) *Stochastic Models, Estimation and Control*. (Vols 1, 2). Academic Press, New York.

Migon, H.S. & Harrison, P.J. (1985) An application of non-linear Bayesian forecasting to television advertising. In *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith (Eds.). North-Holland, Amsterdam and Valencia Press.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. and Verma, T.S. (1987) The logic of representing dependencies by directed graphs. In *Proc., 6th Natl. Conf. on AI (AAAI-87)*, Seattle, 374–379.

Pole, A. and West, M. (1990) Efficient Bayesian learning in non-linear dynamic models. *J. of Forecasting*, 9, 119–136.

Pole, A., West, M. & Harrison, P.J. (1988) Non-normal and non-linear dynamic Bayesian modelling. In *Bayesian Analysis of Time Series and Dynamic Models*. J.C.Spall (Ed.) Marcel Dekker, New York, 167–198.

Priestley, M.B. (1980) State-dependent models: a general approach to non-linear time series analysis. *J. of Time Series Analysis*. 1 (1), 47–71.

Quintana, J.M. (1985) A dynamic linear matrix-variate regression model. *Research Report 83*, Department of Statistics, University of Warwick.

Quintana, J.M. (1987) *Multivariate Bayesian Forecasting Models*. Unpublished PhD thesis, University of Warwick.

Quintana, J.M. and West, M. (1987) Multivariate time series analysis: new techniques applied to international exchange rate data. *The Statistician*, 36, 275–281.

Quintana, J.M. & West M. (1988) Time series analysis of compositional data. In *Bayesian Statistics 3*. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.), Oxford University Press.

Roberts, A. & Docker, J. (1986) Lifestyle or brandstyle: Which is the better to segment by? *ADMAP*, Feb., 78-85.

Robinson, P.M. (1973) Generalised canonical analysis for time series. *J. Multiv. Anal.*, 3, 141-160.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63, 581-592.

Samli, A.C. (1989) Heterogeneity of markets and segmentation in retailing. In *Retail Marketing Strategy*, A.C. Samli (Ed.). New York, Quorum, 135-153.

Shachter, R.D. (1986) Intelligent probabilistic inference. In *Uncertainty and Artificial Intelligence*. L.N. Kanal & J. Lemmer (Eds.) North-Holland, Amsterdam, 371-382.

Smith, A.F.M. & West, M. (1983) Monitoring renal transplants: an application of the multi-process Kalman filter, *Biometrics*, 39, 867-878.

Smith, J.Q. (1979) A generalisation of the Bayesian steady forecasting model. *J. Royal Statist. Soc., B*, 41, 375-387.

Smith, J.Q. (1981) The multiparameter steady model. *J. R. Statist. Soc. B*. 43 No 2 255-260.

Smith, J.Q. (1988) Models, optimal decisions and influence diagrams. In *Bayesian Statistics 3*. J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith (Eds.) Oxford University Press.

Smith, J.Q. (1989) Influence diagrams for statistical modelling. *Annals of Statistics*, 17 (2), 654-672.

Smith, J.Q. (1990) Statistical principles on graphs. In *Influence Diagrams, Belief Nets and Decision Analysis*. R.M. Oliver & J.Q. Smith (Eds.) John Wiley

and Sons Ltd, 89–120.

Smith, J.Q. (1990) Non-linear state space models with partially specified distributions on states, *J. of Forecasting*, **9**, 137–149.

Smith, J.Q. & Young, S.C. (1988) Stochastic experimental games — a Bayesian approach. *Research Report 128*, Department of Statistics, University of Warwick.

Smith, W. (1956) Product differentiation and market segmentation as alternative marketing strategies. *J. of Marketing*, **21**, 3–8.

Spiegelhalter, D.J. & Lauritzen, S.L. (1990) Sequential Updating of Conditional Probabilities on Directed Graphical Structures. *Networks*, **20**, 579–605.

Stock, J.H. & Watson, M.W. (1988) Testing for common trends. *J. Amer. Statist. Assoc.*, **83**, 1097–1107.

Tarjan, R.E. & Yannakakis, M. (1984) Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs. *SIAM J. Comput.*, **13**, 566–579.

Verma, T.S. & Pearl, J. (1990) Equivalence and synthesis of causal models. In *Proceedings, Sixth Conference on Uncertainty in AI, Cambridge, Mass.*, 220–227.

Wermuth, N. & Lauritzen, S.L. (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. R. Statist. Soc. B.*, **52** (1), 21–50.

West, M. (1982) *Aspects of Recursive Bayesian Estimation*. Unpublished PhD thesis, University of Nottingham.

West, M. (1986) Bayesian model monitoring. *J. Roy. Statist. Soc., B*, **48**, 70–78.

West, M. & Harrison, P.J. (1986) Monitoring and adaptation in Bayesian forecasting models. *J. Amer. Statist. Ass.*, **81**, 741–750.

West, M. & Harrison, P.J. (1989a) *Bayesian Forecasting and Dynamic Models*.

Springer-Verlag.

West, M. & Harrison, P.J. (1989b) Subjective intervention in formal models. *J. of Forecasting*, 8, 33–53.

West, M., Harrison, P.J. and Migon, H.J. (1985) Dynamic generalised linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Ass.*, 80, 73–97.

Wind, Y. (1978) Issues and advances in segmentation research. *J. of Marketing Research*, 15, 317–337.

Young, S.C. & Smith, J.Q. (1991a) Deriving and analysing optimal strategies in Bayesian models of games. *Management Science*, (to appear).

Young, S.C. & Smith, J.Q. (1991b) Suboptimality of M-step back strategies in Bayesian games. *J. of Game Theory*, (to appear).

Zeeman, E.C. (1977) *Catastrophe Theory : Selected Papers (1972–1977)* Addison-Wesley.

Zellner, A. (1986) *Basic Issues in Econometrics*. The University of Chicago Press.